Supplementary materials for Stark, Lin, Kheradpour, Pedersen *et al.* (2007)

Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures

Table of Contents

S1 Alignments and phylogeny

We generated three different sets of whole-genome sequence alignments (of *D. melanogaster* with the 11 other flies) for use in various parts of this study. Two were derived from a synteny map generated by MERCATOR (C. Dewey, `http://www.biostat.wisc.edu/~cdewey/mercator/`), with sequence alignments generated by MAVID[1] and PECAN (B. Paten and E. Birney, `http://www.ebi.ac.uk/~bjp/pecan/`). Additionally, we generated MULTIZ[2] alignments of 15 insects, ignoring the non-*Drosophila* species in that alignment.

We estimated branch lengths in the phylogenetic tree for the flies (shown in Figure 1 and referred to throughout) based on four fold degenerate sites in alignments of orthologous protein-coding genes. We identified one-to-one orthologs based on based on FlyBase annotation release 4.3 for *D. melanogaster* and community annotations for the 11 other species (*Drosophila* Sequencing and Analysis Consortium, *Nature*, submitted), yielding 12,861 four fold sites. Then, to estimate branch lengths, we ran PHYML[3] v2.4.4 with an HKY model of sequence evolution, a fixed tree topology (Figure 1a), and remaining parameters at default values. For comparison with vertebrates in Figure 1, we estimated the branch lengths for 28 vertebrates using 10,340 four fold sites, based on alignments of genes with one-to-one orthologs in human, dog, and mouse (M. Clamp, pers. comm.). The alignments of the fly orthologs were reverse translated from MUSCLE[4] peptide alignments, while the vertebrate alignments were extracted from MULTIZ alignments downloaded from the UCSC Genome Browser[5].

S2 Gene identification

S2a Supplementary Methods

RFC and CSF metrics

The **Reading Frame Conservation (RFC)** metric was applied as previously described[6,7]. Briefly, given an alignment of a region of the target genome (*D. melanogaster*), a pairwise score between the target and each informant is computed as the percentage of target nucleotides that align in the same reading frame in the informant (taking the largest such percentage out of the three possible reading frames), where a nucleotide might not occur in the same reading frame due to an upstream frame-shifting indel. Each informant then votes +1, -1, or 0 based on an informant-specific cutoff on the pairwise RFC score: +1 if the score is above, -1 if the score is below, or 0 if there was no sequence aligned. These votes are then summed to obtain an overall score for the region. The cutoff for each species is chosen by examining the usually bimodal distribution of the score between known coding and non-coding regions, and typically ranges between 70-80%.

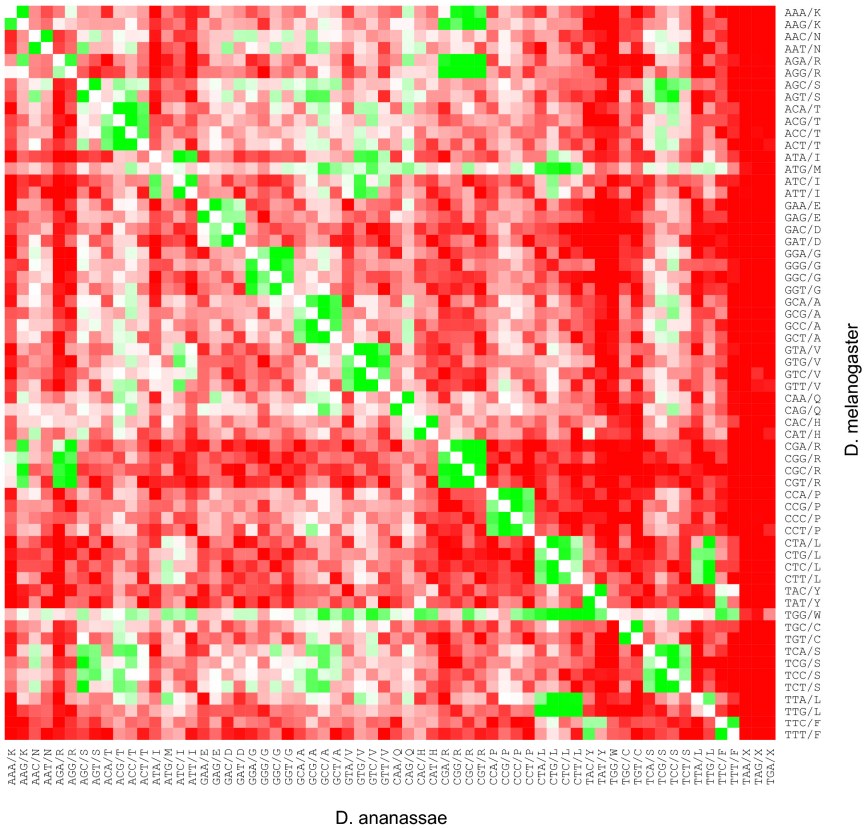
The **Codon Substitution Frequencies (CSF)** metric is based on estimates of the frequencies at which all pairs of codons are substituted between genes in the target species and the informants. First, let us consider computing the score for a pairwise alignment only. Consider the alignment of a putative ORF/exon as two sequences of codons A and B, where A_k is the target codon that aligns to the informant codon B_k at position k in the target codon sequence (position $3k$ in the in-frame target

nucleotide sequence). CSF assigns a score to each codon position *k* where: (1) A_k and B_k are both un-gapped triplets, (2) A_k is not a stop codon, and (3) $A_k \neq B_k$. CSF then sums these scores to obtain an overall score for the sequence.

The score assigned to a codon substitution (*a*,*b*) is a log-likelihood ratio indicating how much more frequently that substitution occurs in coding regions than in non-coding regions. Each likelihood compared in this ratio is derived from a Codon Substitution Matrix (CSM), where

$$CSM_{a,b} = P(\text{informant codon b} \mid \text{target codon a}, a \neq b)$$

The entries of the CSM are estimated for each target and informant by counting aligned codon pairs in training data, and then normalizing the rows to obtain the desired conditional probabilities. We train two CSMs, one for which the training data is alignments of known genes (CSM^C) and one for which the training data is alignments of random non-coding regions (CSM^N). The score that CSF assigns a codon substitution (*a*,*b*) is then $\log \frac{CSM^C_{a,b}}{CSM^N_{a,b}}$. For example, these scores for *D. melanogaster* and *D. ananassae* are shown in the following visualization, in which green represents a positive score and red represents a negative score:



With multiple informants, CSF uses an *ad hoc* strategy to combine evidence from the informants without double-counting multiple apparent substitutions among extant species that result from fewer evolutionary events in their ancestors. For each target codon position *k*, CSF assigns a score to codon substitutions between the target and each informant exactly as in the pairwise case, using the appropriate CSMs for each informant. CSF then takes the median of these scores to obtain a composite score for position *k*, and sums these composite scores to obtain an overall score for the sequence. Note that the median is usually taken on fewer than *n* pairwise scores, since the pairwise scores are only assigned to ungapped informant codons that differ from the target codon.

Lastly, we note that CSF makes no attempt to explicitly "correct" for several well-known issues that frequently arise in modeling codon evolution, such as transitions/transversions, CpG hypermutation, codon bias, site-specific rate variation, etc. The purpose of CSF is neither to realistically model evolution nor to obtain precise estimates of evolutionary rates, but rather to provide a computationally efficient metric that discriminates between coding and non-coding regions. In addition to the controls and benchmarks in this paper, we have verified its effectiveness for this purpose by cross-validated benchmarks in comparison to several other methods[8].

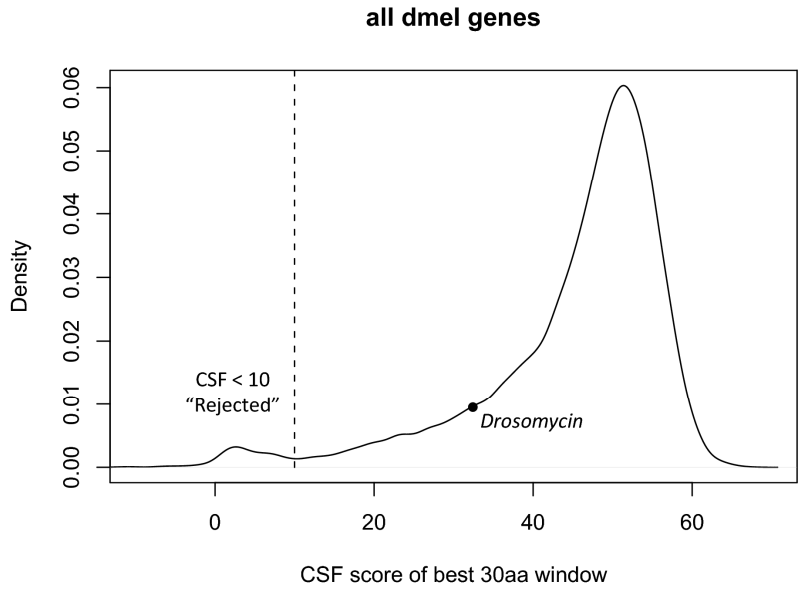
Confirming and rejecting existing gene annotations

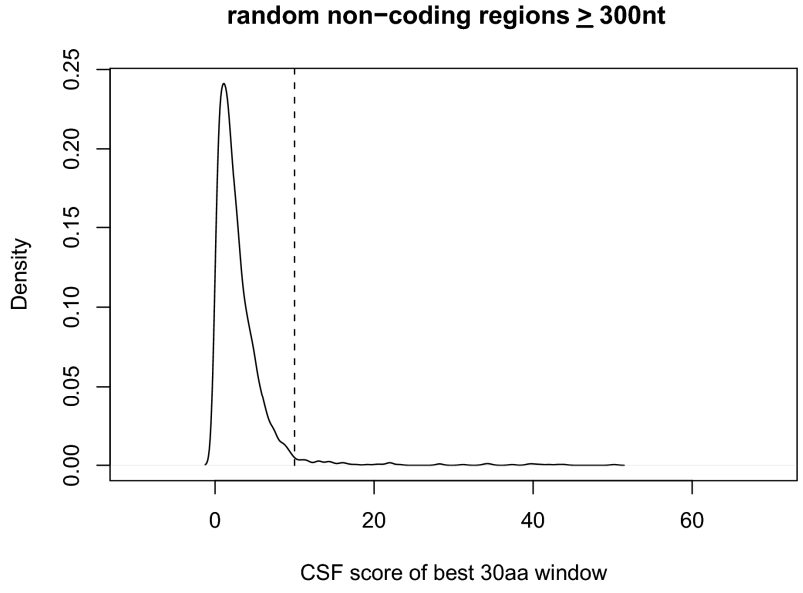
For each euchromatic gene in FlyBase annotation release 4.3, we applied the RFC and CSF metrics to each of its transcript models. To score a transcript, we first generated an alignment by extracting each of its exons from whole-genome sequence alignments and then "splicing" them. We then used the best-scoring transcript model as a proxy for the gene, where the best-scoring transcript model is the one with the highest RFC score, or, in the event of a tie of the RFC score, the highest CSF score.

To define a test for whether the evolutionary evidence "confirms" each gene, we chose cutoffs on the RFC and CSF scores based on random controls as follows. We extracted 15,564 regions ≥300nt in length from the genome sequence alignments, chosen uniformly at random from the portion of the genome not annotated as protein-coding. These alignments were preprocessed to remove columns containing in-frame stop codons in *D. melanogaster* (each control region is ≥300nt in length *after* removing stop codons, a detail previously omitted) and then scored by RFC and CSF. We considered a gene "confirmed" if its RFC score was greater than zero and its length-normalized CSF score (the CSF score divided by the length in nucleotides of the ORF) was greater than or equal to 0.03, cutoffs which exclude all but three of the 15,564 control regions (see Table 2). One of these three "false positive" regions coincided with a predicted new exon that was later validated by our cDNA sequencing experiments, and, following manual inspection, we consider the other two also likely to represent genuine coding sequence. Thus, our criteria for "confirmation" of a gene was very stringent, insofar as virtually no non-coding regions ≥300nt passed this test.

We next defined a much more relaxed test to identify gene annotations that not only fail to satisfy the above stringent criteria, but appear unlikely even to represent genuine protein-coding genes. We computed the CSF score over every overlapping 30aa window in every transcript model for each gene. Additionally, we computed these scores using the three different genome alignment sets and using

three different subsets of the informant species, representing all twelve *Drosophila* genomes, the subgenus *Sophophora*, and the *melanogaster* group. We took the highest scoring window in each gene, out of all its transcripts, all of the alignments, and all of the phylogenetic clades, as the score for that gene. The distribution of this score across all genes was clearly bimodal:





Also shown on the above plot is the score of *Drosomycin*, a recently evolved gene unique to the *melanogaster* group. We chose a cutoff selecting the 454 genes forming the lower distribution as the "rejected" genes. Genes that were neither "confirmed" nor "rejected" by these tests form the "abstain" category (Table 2).

The manual review of the "rejected" genes (as well as the predicted news exons, below) was carried out according to FlyBase Gene Model Annotation Guidelines, described at:

http://flybase.bio.indiana.edu/static_pages/docs/refman/refman-G.html#G7

Predicting new exons

To define the precise boundaries of genomic regions showing RFC and CSF evolutionary signatures that are likely to represent new exons, we integrated our evolutionary metrics as features into a simple *de novo* exon predictor based on a semi-Markov conditional random field[9,10]. (SMCRF), a probabilistic graphical model similar to a generalized hidden Markov model (GHMM). Unlike a GHMM, however, an SMCRF can *directly* incorporate any metric that provides a real-valued score for any segment of the genome, such as RFC and CSF. Our system is a straightforward application of standard SMCRF algorithms to parse the genome into coding and non-coding segments based on our metrics. In this sense, it may be considered more similar to simple interval segmentation algorithms that compute boundaries of high-scoring regions, than to full gene predictors such as GENSCAN[11] or N-SCAN[12]. Initial applications of SMCRFs to create full gene predictors have recently been reported[13,14].

SMCRF structure. The graphical structure (state diagram) of our model follows the example of ExoniPhy[15], with some simplifications enabled by the more flexible nature of the SMCRF than the phylo-HMM used in that system. In particular, the model has only seven segment labels (states): one for each codon reading frame on each strand (+1, +2, +3, -1, -2, -3), and one for non-coding positions. Since each coding state labels a segment, not an individual nucleotide, the labels (+1, +2, +3, -1, -2, -3) specify the codon reading frame in which the segment should be read. For example, the label +1 means that the segment should be read as beginning on the first position of a complete codon on the positive strand. Each "coding" state is bidirectionally connected to the non-coding state, the non-coding state is self-connected, and no other transitions are possible.

If there is no maximum segment length, then the SMCRF training and decoding algorithms have running time quadratic in the sequence length. Therefore, for practical reasons, non-coding "segments" are constrained to be one nucleotide in length, with non-coding regions modeled as sequences of 1nt non-coding segments. The maximum length of coding segments is *de facto* constrained by disallowing in-frame stop codons.

Feature functions. The features used by the SMCRF include:

1. the evolutionary metrics, which score coding segments.

2. indicator functions for start and stop codons, which score transitions between coding and non-coding segments. These are binary functions (later assigned a numerical weight by the SMCRF) indicating the presence of a start or stop codon in the *D. melanogaster* sequence. They also enforce "well-formedness" constraints: the start codon feature disallows (by returning a negative-infinity score) noncoding-to-coding transitions in the absence of a start codon or AG

splice site, and the stop codon feature disallows coding-to-noncoding transitions in the absence of a stop codon or GT splice site. The stop codon feature also disallows coding segments with in-frame stop codons.

3. sequence-based discriminators for acceptor and donor sites, which score transitions between coding and non-coding segments on AG and GT splice sites (based on the *D. melanogaster* sequence). These discriminators, provided by the authors of a previous study[16], consider 23 nucleotides surrounding acceptor sites and 9 nucleotides surrounding donor sites based on the principle of maximum entropy.

4. length distribution feature, which was set to a simple geometric distribution corresponding to the empirical mean lengths of annotated exons and non-coding regions. (We did not investigate other exon length distributions at the time of freezing our prediction set for this study, although this is possible in principle.)

Importantly, the SMCRF did not include any explicit coding sequence composition features (e.g. high-order Markov models), nor did it use any information about transcript sequence evidence or homology to known proteins.

Training and decoding. The SMCRF training procedure determines optimal weights for a linear combination of the features. We trained our SMCRF using the standard maximum conditional likelihood algorithm[9,10] on a training set of 100 known genes. (The SMCRF training procedure must estimate only one parameter for each feature, in contrast to GHMM gene predictors which require thousands of generative parameters. The SMCRF thus requires less training data. In our case, much of the additional information that would be estimated in GHMM training is captured in the CSMs used by CSF.) We then used the SMCRF equivalent of the Viterbi algorithm to decode the whole *D. melanogaster* genome in the Mercator/MAVID alignments into coding exons and non-coding regions.

All predicted exons that did not overlap any coding exon in FlyBase annotation release 4.3 (on the coding strand) were regarded as predicted new exons, except for a total of 217 predictions that were either within *Dscam* (a gene with exceptionally many exons and splice forms that are known but not all represented in FlyBase), heterochromatic regions, or redundant or misassembled regions of the euchromatic genome assembly. For historical reasons, the new exon predictions were carried out only using the Mercator/MAVID alignments; the MULTIZ alignments of the 12 flies were not available until our experimental validation and manual curation efforts were already underway for a frozen prediction set (see also Figure S8).

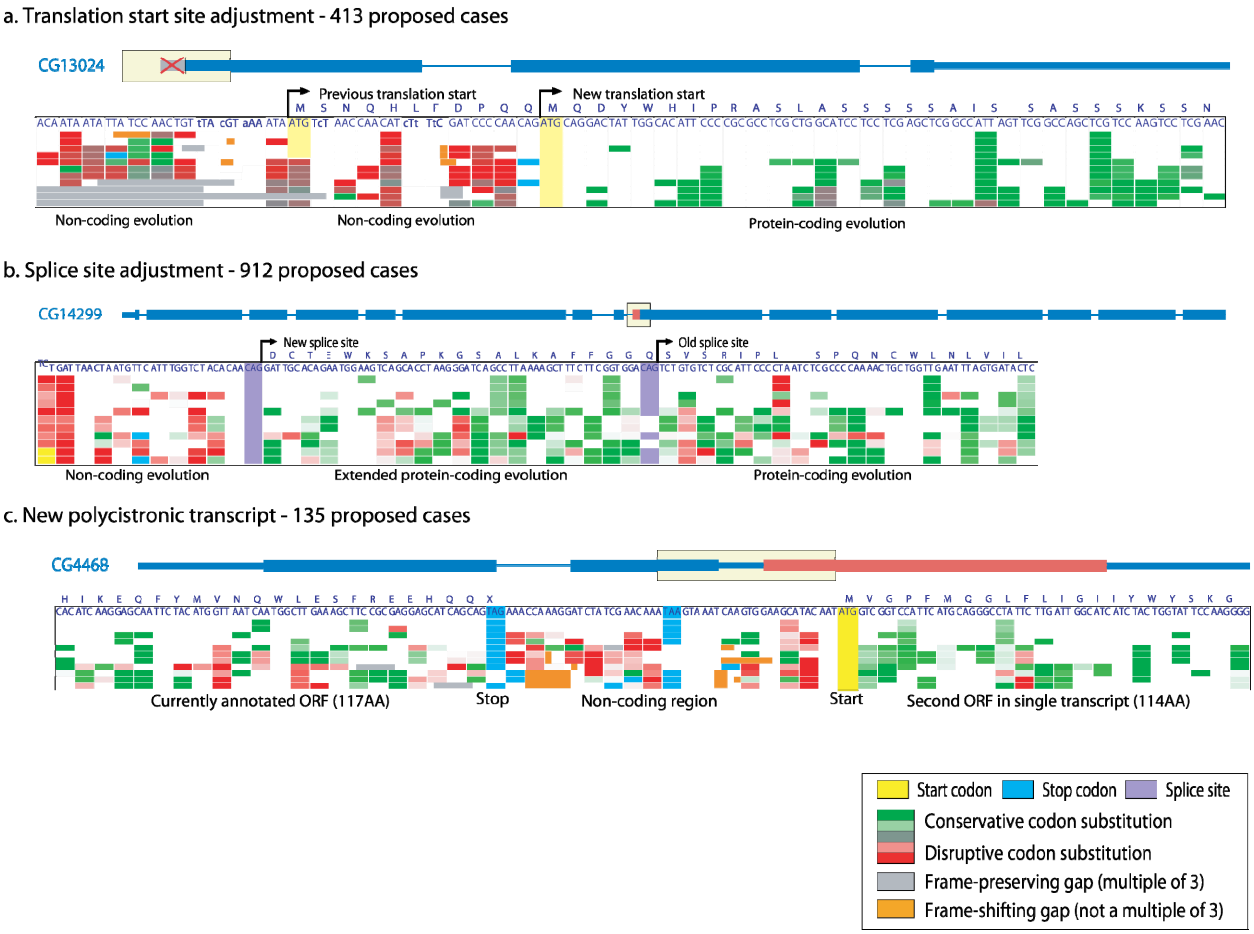


Figure S2b. Additional examples of adjustments to existing protein-coding gene annotations and unusual gene structures. (a) Proposed adjustment to translation start site. The longest complete ORF in this transcript has been annotated as protein-coding, but evolutionary signatures suggest that translation starts at a downstream ATG. (b) Comparative evidence supports a 21aa extension of an annotated protein-coding exon. (c) Protein-coding evolution of an ORF within the annotated UTR of an existing transcript model suggests a dicistronic transcript structure.

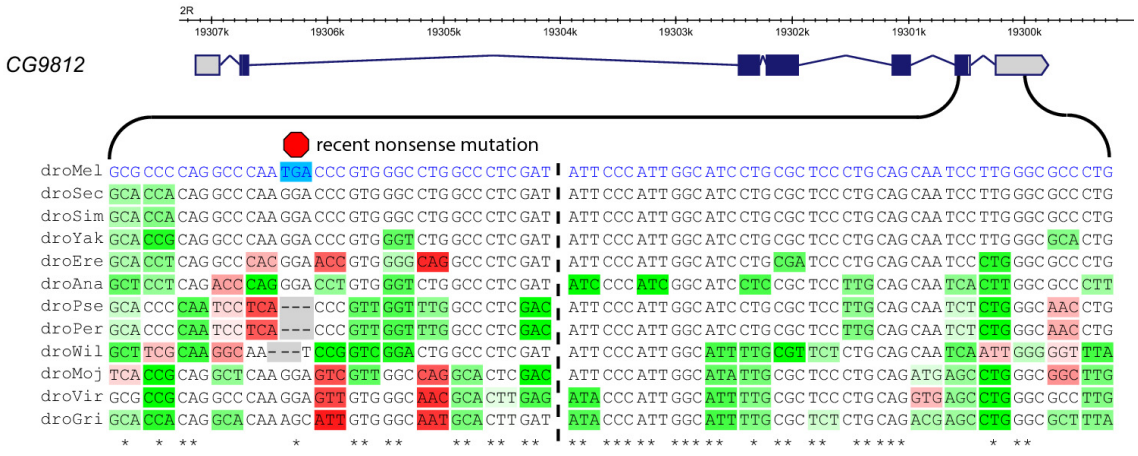


Figure S2c. Example of a recent nonsense mutation in *D. melanogaster* identified by evolutionary signatures. The signatures clearly extend past the stop codon, which is not present in the informants. In contrast, the predicted stop codon readthrough genes have the stop codon conserved across the informant species.

S3: RNA structure and gene prediction: Supplementary Methods

The genomic screen based on EvoFold[17], the definition of high-confidence prediction sets, and the subsequent detailed analysis of these sets are described below.

Genomic screen of structural RNAs

Since paired regions of structural RNAs evolve slowly, we focused our screen on the conserved segments of the 12-way *Drosopholid* multiz alignment (Methods S1). Conserved elements were identified using phastCons[18] and subsequently extended by 20 bases and combined when overlapping to also include fast-evolving single-standed regions. These conserved elements contain 98% of all known noncoding RNAs while spanning 58% of the *D. melanogaster* genome. EvoFold was then applied to both strands of each corresponding alignment segment in overlapping windows of length 240 with offset of 80.

Weak predictions, with less than ten base pairs or more than 50% bulges in stems, were removed together with folds overlapping simple/low-complexity repeats (see definitions below). Finally, a single coverage set was extracted by retaining only the highest scoring folds when overlap occurred. This resulted in the initial comprehensive set of 22,682 predictions.

High confidence prediction sets

Since a high fraction of false positives are inherent in comprehensive genomic screens of structural RNAs[19], high-confidence prediction sets were constructed for detailed analysis. Because of the distinct evolutionary signature of structures overlapping protein-coding regions, one set was defined for intergenic, intronic, and untranslated regions and a separate set for protein-coding regions.

The high-confidence set for intergenic, intronic, and untranslated regions consists of predictions with strong substitution evidence, including only predictions with at least two compensatory changes and at least one compensatory change for every two disruptive changes. Since alignment quality influences the substitution counts strongly, e.g., a wrongly aligned sequence may cause an otherwise compelling candidate to be excluded, poorly aligning sequences were removed before the substitution counts were made. A sequence was removed if: (1) it contains more than 5% annotated repeats, (2) more than 7.5% of the positions with predicted pairs are gapped, or (3) if more than 10% of the predicted pairs contain contradictory substitutions and all other sequences are perfectly conserved. If the resulting alignment contains less than 5 sequences, it is removed in its entirety.

Structures overlapping protein-coding regions are generally deeply conserved and experience only few substitutions due to the evolutionary constraints at both the protein level and RNA structure level. The majority of the known structures of this type are RNA editing substrates[20]or other long hairpins[21,22]. Since the EvoFold log-odds score rank these highly, we define the high-confidence protein-coding prediction set as top-50% percent hairpins with more than 15 pairing bases.

Enrichment and significance statistics

Structure prediction enrichments: Functional classes of genes are defined using the terms of the Gene-Ontology database[23] and gene locations are defined using the UCSC Browser set of 13566 “Canonical

Transcripts" (see definition below). P-values for enrichment of predictions for a given functional class are based on the hyper-geometric distribution and calculated with the R program (http://www.R-project.org).

Transcription evidence enrichment: The transfrags from 12 tiling array experiments were combined to a single genomic map of transcribed regions[24]. The intergenic structure predictions were then randomly placed within the intergenic regions and the overlap with the transcribed regions measured. This procedure was repeated 1000 times. The fraction of times the measured overlap was larger or equal to the overlap originally observed is reported as the P-value of enrichment.

Strand-bias calculation: Because of the strand symmetry of A-U (U-A) and G-C (C-G) base pairs, many RNA structures appear supported by compensatory substitutions on both DNA strands. In contrast, the reverse complement of G-U (U-G) is A-C (C-A), which cannot base pair. Substitutions involving G-U (U-G) pairs leads to a small strand-dependent difference in the EvoFold score for true RNA structures[17]. When overlapping predictions occur, the highest scoring prediction is chosen and a strand is thereby assigned. Cis-regulatory RNA structures of protein-coding genes will be found on the transcribed DNA strand. For the transcribed regions of the genome expected to contain cis-regulatory structures, we therefore calculate a strand bias, defined as the fraction of structures found on the transcribed strand. For RNA structures overlapping protein-coding regions, which evolve extremely slowly with few substitutions, the EvoFold score does not allow efficient strand assignment.

Scaling

To evaluate how discovery power scales with the number of species and the branch lengths of the input alignment (see main text and section S7b), we study our ability to recover known ncRNAs in different subsets of the 12-way alignment. Starting with the 12-way MULTIZ alignment, subset multiple alignments were created by sequentially removing the species most distant to DM. Pairwise alignments referenced by DM were likewise extracted from the 12-way MULTIZ alignments. For each of these alignments, we ranked predictions in the comprehensive prediction set by their substitution score ([2*#compensatory - #contradictory]/#pairs), and counted the number of known ncRNAs in top-100.

Data sets

The various data sets used to annotate and categorize the prediction sets are defined below.

Gene set and genomic regions: The FlyBase version 4.3 *Drosophila melanogaster* gene definitions were used throughout. Predictions were assigned to genomic regions based on this set. If a prediction overlap the boundary of two types of genomic regions, preference was given as follows: protein-coding > 5'UTR > 3'UTR > intronic > intergenic.

Canonical transcript set: The set of "Canonical Transcripts" defines a single representative transcript for each gene. It can be downloaded from the UCSC Genome Browser SQL-database[25].

Repeats: Repeats were defined by the "repeats" track of the UCSC Genome Browser, which is made with the repeat-masker program. Simple and low-complexity repeats were extracted from this set.

Known structural RNAs: The reference set of known ncRNAs was defined as the non-coding RNAs genes of FlyBase verson 4.3. For consistency with the miRNA analysis, the miRNAs of this set were replaced with validated miRNAs from mirBase release 9.0[26]. The final set contains a total of 613 ncRNAs.

Two known structural RNAs (a SECIS and a TLS element) were identified among the 38 high confidence 3'UTR structural RNAs by searching the literature as well as relevant databases (rfam and UTRdb)[27-29].

Known A-to-I RNA editing events: The reference set of known A-to-I editing event was defined as the validated cases reported in[20,30-32].

Gene ontology: The May 2006 version of the Gene ontology (GO) database was used[23].

S4 miRNA gene prediction

S4a. Supplementary Methods

miRNA training sets

From miRBase release 9.0[26], we selected all 60 *Drosophila melanogaster* miRNAs that have been cloned. For both hairpin and mature prediction, miRNAs that may bias the score due to overfitting are excluded when scoring the known miRNAs (see the appropriate section below for details).

Collecting all melanogaster hairpins

To identify miRNA-like hairpins, we ran RNAfold from the Vienna package[33] on 120 nt windows (overlap of 90 nts) in the *Drosophila* melanogaster genome (rel 4). We considered all hairpins in each window (including branching hairpins) and trimmed them to the end of the stem. We use these folds to infer the arms and loop of each hairpin. As a lenient prescreening, we removed all hairpins shorter than 63 nts, with an arm of less than 20 nts or with less than 70% arm base-pairing. We were left with all the known hairpins and an additional 760,000 potentially overlapping list of putative miRNA hairpins.

Hairpin sequence alignments

For each melanogaster hairpin sequence, we selected the best BLAST[34] match with E-value ≤ $1x10^{-5}$ in each of the 11 other genomes (CAF1 assemblies). We performed a multiple alignment of the corresponding sequences plus 50 nt flanking sequence on each side using ClustalW[35].

miRNA hairpin discovery

For each hairpin we derived several structural and conservation features. These features are listed Supplemental Table S4b. We scored the list of 760,355 putative miRNAs with a method similar to Random Forests[36]. Using the combined conservation and structural feature set, 500 decision trees were trained on the positive training set of 60 miRNAs and a different randomly selected negative set of 250 of the remaining putative miRNAs. The final score for a hairpin was derived through cross validation (the score of each hairpin is evaluated only by trees that exclude it, all redundant sequence similar miRNAs and overlapping miRNAs). From all hairpins that overlap on the same strand, only the hairpin with the highest score is kept. This can lead to known miRNAs having a slightly revised hairpin selected.

miRNA mature 5' identification

For each position in the hairpin, we computed several features indicative of the start (5'end) of mature miRNAs. 7mer scores are determined for the sequence complementary 7mers for each position in the hairpin. 7mer conservation scores are motif-conservation scores (MCS) of 7mers calculated in all annotated 3'UTRs (FlyBase release 4.3) as described in[6,7,37]. Additionally, we assessed the avoidance of the 7mers in 3'UTRs of global anti-target genes[38] by computing the deviation relative to all genes by Z-scores. In addition, we considered the nucleotide to account for the U-bias often observed in mature miRNAs, the number of paired bases in a window of 7 around the position, and others (see Supplemental Table S4c). We excluded potential start positions for which the corresponding miRNA would fall outside the hairpin or span the hairpin loop region (for positions in the left arm, we required at least 15 nts before the start of the loop; for positions in the right arm, we allowed no more than a 3 nt

overlap with the loop and required at least 18 nts before the end of the hairpin). Within each hairpin, we linearly normalized each feature to be from 0 to 1 and marked each known mature site as a positive and all remaining sites as negative. We augmented the features for each position with the features of the position of the left and the right. We used the SVMlight[39] package to train an SVM with default parameters (linear kernel and positive gain 1) on all the permissible locations from all the known hairpins. The SVM scores for each hairpin are linearly normalized so that the scores of the permissible regions have mean 0 and standard deviation 1. We predict the mature location by taking the permissible location in each hairpin with the highest SVM score. Each hairpin is only scored by models trained on cloned Rfam hairpins (Rfam 9.0), excluding itself and all family members. For evaluation (not training), we use the partly corrected 5'end annotation of[40]. To test if we predicted the star sequences, we determined the star sequence based on the fold-back structure as a 2nt 5'overhang of the mature miRNA sequence.

Validation of novel miRNAs

To validate our predictions experimentally, we obtained 763,111 Solexa reads corresponding to 1524 distinct sequences that matched to our predictions. These were cloned from adult *Drosophila* ovaries and testes as described previously[41]. We excluded short reads (<15 nucleotides) and those that matched the genome more than 3 times and aligned the remaining reads to the predicted hairpins. For validation, we required that at least one position in the hairpin was supported by at least 10 reads and manually inspected the alignments for miRNA-like processing patterns (e.g. dominant sequence, presence of star sequence, no sign of degradation products). In addition, we intersected the predicted hairpins with curated sequencing reads of several *Drosophila* libraries kindly provided by Graham Ruby, David Bartel and Eric Lai. To validate mature miRNA 5'end predictions, we used the curated mature miRNAs reported by[40], and refer to them by their newly assigned Rfam names[26].

miRNA recovery using different species sets

We investigated the dependency of genome-wide miRNA discovery on the number and evolutionary distance of the contributing species. For this, we obtained all novel miRNAs defined by[40], and tested how many we recovered with our protocol when using selected subsets of species. In each case, we allowed for the optimal re-weighting of feature contributions (e.g. to down-weight conservation features when comparing only close species, if appropriate).

Supplementary Table S4b: Features for hairpin prediction and their performance

	Range (5% - 95%)			Known enrichment
Feature	Known miRNAs	miRNA-like hairpins	Random hairpins	
Correlation to average conservation profile of known miRNAs	0.75 - 0.97	-0.28 - 0.7	-0.28 - 0.77	327.54
Match fraction for the less conserved arm	5.9 - 12	1.8 - 8.9	1.6 - 8.1	42.32
Z-score of the minimum free energy (MFE) with respect to the length and GC content	2.7 - 6.2	0.11 - 3.4	-2.5 - 2.5	39.25
Match fraction for the more conserved arm	6.1 - 12	2.5 - 10	2.5 - 10	16.14
Mismatch fraction for the region flanking the more conserved arm	1.9 - 6.4	0.05 - 3.7	0.05 - 3.7	13.38
Mismatch fraction for the loop	0.33 - 5.5	0 - 2.2	0 - 2.1	10.87
Number genomes the hairpin is present in (range 2..12)	7 - 12	3 - 12	3 - 12	9.56
Mismatch fraction for the region flanking the less conserved arm	2.4 - 6.2	0.1 - 4.6	0.09 - 4.7	6.56
MFE of the consensus fold	-41 - -20	-30 - -2.5	-27 - -1.6	6.32
Average difference between the MFE of the consensus fold and the individual fold	0.59 - 9.4	1.9 - 22	1.3 - 21	6.24
(less conserved arm conservation)/(loop conservation)	0.99 - 1.7	0.63 - 1.2	0.53 - 1.2	5.18
Number of bases conserved in the longest stretch of perfectly conserved base pairs in an arm	20 - 39	5 - 29	1 - 22	5.00
Match fraction for the region flanking the less conserved arm	3.9 - 8.6	1.7 - 7.9	1.7 - 7.6	4.68
(more conserved arm conservation)/(loop conservation)	1 - 1.8	0.86 - 1.4	0.87 - 1.3	4.11
Match fraction for the loop	4.1 - 12	2 - 10	2.2 - 9.7	3.94
(MFE of consensus fold/average MFE of the individual ortholog folds)	0.68 - 1	0.15 - 0.96	0.12 - 0.97	3.74
Match fraction for the region flanking the more conserved arm	3.9 - 8.7	2 - 8.8	1.9 - 8.6	3.37
Mismatch fraction in less conserved arm	0.036 - 1.8	0.19 - 3.4	0 - 4.4	3.28
(# of symmetric bulges)/(# of asymmetric bulges)	0.33 - 7	0.000025 - 3	0.00005 - 10000	3.01
Mismatch fraction in more conserved arm	0 - 1.3	0 - 1.4	0 - 1.8	2.60
Number of paired bases in the best stretch of 22 base pairs	18 - 22	16 - 21	3 - 19	2.56

Indel fraction for the loop	0 - 0.16	0 - 0.17	0 - 0.14	2.44
Percent of arm bases that are paired	0.77 - 0.9	0.72 - 0.86	0.67 - 1	2.32
Number of bases that need to be removed to make all internal bulges symmetric	1 - 7	1 - 13	0 - 10	1.99
Indel fraction for the region flanking more conserved arm	0.012 - 0.17	0 - 0.14	0 - 0.15	1.97
Less conserved arm length	31 - 47	21 - 52	3 - 42	1.72
More conserved arm length	31 - 47	21 - 52	3 - 42	1.72
Indel fraction for the less conserved arm	0 - 0.036	0 - 0.23	0 - 0.26	1.66
Indel fraction for the region flanking less conserved arm	0.018 - 0.17	0 - 0.2	0 - 0.2	1.49
Loop length	4 - 26	4 - 64	4 - 110	1.37
Structure length	73 - 100	65 - 120	47 - 120	1.36
Number of substructures	0 - 0	0 - 1	0 - 2	1.36
Indel fraction for the more conserved arm	0 - 0.02	0 - 0.078	0 - 0.07	1.21
Number of internal bulges	3 - 8	2 - 9	0 - 6	1.20

Supplementary Table S4c: Features for mature miRNA prediction and their performance

	Range (5% - 95%) or Faction 1 (for binary features)		Known Enrichment
Feature	Known 5'	Other positions	
Target 7mer conservation profile correlation	-0.12 - 0.94	-0.49 - 0.65	15.2
Target 7mer conservation score	5.9 - 110	-0.54 - 24	11.04
Target 7mer avoidance score	-1.5 - 4.6	-3.5 - 3.7	4.11
Length of perfect conservation that follows, including start	15 - 32	0 - 32	3.54
20mer conservation	0.99 - 1	0.83 - 1	2.85
0/1 presence of U	0.78	0.30	2.62
0/1 start of perfect 20mer conservation	0.95	0.42	2.24
0/1 start of perfectly conserved 8-mer followed by a 95%-conserved 12-mer	0.97	0.45	2.15
Size of symmetric loop	0 - 2	0 - 2	1.9
Distance from terminal loop	0 - 9	-5 - 20	1.46
Distance from start of the hairpin	3 - 16	-2 - 23	1.43
Overlap length with the loop region	0 - 0	0 - 5	1.24
0/1 loop at 1 position	0.27	0.22	1.21
Number of paired bases in window of 7	3 - 7	2 - 7	1.18
0/1 loop at 2 position	0.08	0.21	1.16
0/1 loop at -2 position	0.18	0.26	1.11
Size of overlapping bulged loop	0 - 0	0 - 6	1.08
0/1 loop at -1 position	0.22	0.25	1.04
Number of paired bases in window of 3	1 - 3	0 - 3	1.03
Number of paired bases in window of 5	2 - 5	1 - 5	1.02
0/1 loop at current position	0.23	0.23	1

Supplemental Table S4d: Known and predicted miRNAs

Score	Val	Locus	Name	MatureChange	Host Gene	Species	Targets	Mature	Orthologs
1.000	HM	2L 11953414 11953499 -	miR-263a	11->14		12+	129	AATGGCACTGGAAGAATTCACGGG	hc
1.000	HM	2L 16693853 16693944 +	miR-9c*		grp	11+	202	TCTTTGGTATTCTAGCTGTAGA	dhca
1.000	HM	2L 16694495 16694567 +	miR-79*		grp	11+	230	TAAAGCTAGATTACCAAAGCAT	ha
1.000	HM	2L 17562309 17562385 +	miR-124*			12+	127	TAAGGCACGCGGTGAATGCCA	hca
1.000	HM	2L 19566106 19566198 -	miR-2b-2*		spi	12+	293	TATCACAGCCAGCTTTGAGGAGCG	dca
1.000	HM	2L 20475034 20475124 +	miR-1*			12+	172	TGGAATGTAAAGAAGTATGGAG	hca
1.000	HM	2L 20604215 20604299 +	miR-133			12+	30	TTGGTCCCCTTCAACCAGCTGT	ha
1.000	HM	2L 7425801 7425886 +	miR-275*			12+	47	TCAGGTACCTGAAGTAGCGCG	da
1.000	L	2L 9950419 9950518 -	miR-87			12+	9	TTGAGCAAAATTTCAGGTGTGTGA	c
1.000	HM	2R 15175570 15175671 -	miR-6-3*			12+	299	TATCACAGTGGCTGTTCTTTTT	dca
1.000	HM	2R 15175862 15175950 -	miR-6-1*			12+	299	TATCACAGTGGCTGTTCTTTTT	dca
1.000	L	2R 15176008 15176098 -	miR-5*			12+	109	AAAGGAACGATCGTTGTGATATG	
1.000	H	2R 15176151 15176233 -	miR-4*			12+	226	ATAAAGCTAGACAACCATTGAA	c
1.000	HM	2R 15176285 15176387 -	miR-286*			12+	118	TGACTAGACCGAACACTCGTGCT	dca
1.000	H	2R 4824837 4824929 -	miR-987			12+	173	TAAAGTAAATAGTCTGGATTGATG	h
1.000	HM	2R 5065561 5065650 +	miR-14*			12+	132	TCAGTCTTTTTCTCTCTCCTAT	ha
1.000	LM	2R 5133062 5133153 -	miR-307*		Mmp2	12+	62	TCACAACCTCCTTGAGTGAGCGA	ca
1.000	H	2R 7674782 7674866 -	miR-988		CG8877	11+	15	CCCCTTGTTGCAAACCTCACGC	
1.000	H	2R 7686061 7686150 -	miR-281-2*	59->58;14	Oda	12+	52	AAGAGAGCTATCCGTCGACAGTC	
1.000	HM	2R 7686280 7686369 -	miR-281-1*		Oda	12+	52	TGTCATGGAATTGCTCTCTTTGT	ca
1.000	HM	3L 10293737 10293827 +	miR-276b*			12+	16	TAGGAACTTAATACCGTGCTCT	d
1.000	H	3L 11920047 11920134 -	miR-285*			11+	39	TAGCACCATTCGAAATCAGTGCT	dhc
1.000	HM	3L 17280267 17280359 +	miR-219			12+	117	TGATTGTCCAAACGCAATTCT	ha
1.000	H	3L 18826231 18826312 +	miR-315*			12+	385	TTTTGATTGTTGCTCAGAAAGCC	a
1.000	H	3L 21507860 21507959 +	miR-193			12+	20	TACTGGCCTACTAAGTCCCAAC	hc
1.000	HM	3L 21602554 21602644 -	miR-316*			12+	178	TGTCTTTTTCCGCTTACTGGCG	
1.000	HM	3L 3234555 3234649 +	miR-282	12->16		12+	27	TAGCCTCTACTAGGCTTTGTCT	
1.000	HM	3L 8545756 8545843 +	miR-190		rhea	12+	116	AGATATGTTTGATATTCTTGGTTG	hc
1.000	H	3R 121093 121174 +	miR-929		cpx	12+	20	CTCCCTAACGGAGTCAGATTG	
1.000	H	3R 21414588 21414680 -	miR-1000		msi	12+	147	ATATTGTCCTGTCACAGCAGT	
1.000	HM	3R 25041316 25041404 +	miR-279*			12+	118	TGACTAGATCCACACTCATTAA	dca
1.000	HM	3R 25042907 25043002 +	miR-996			12+	119	TGACTAGATTTCATGCTCGTCT	dca
1.000	HM	3R 2635223 2635309 -	miR-10*	14->54;14		12+	167	CAAATTCGGTTCTAGAGAGGTTT	
1.000		3R 27091338 27091410 -	Novel-8			12+	479	CAAATTAACTGCGACATGGC	
1.000	HM	3R 5916853 5916932 +	miR-317*			12+	93	TGAACACAGCTGGTGGTATCC	a
1.000	HM	3R 5925744 5925839 +	miR-277*			12+	582	TAAATGCACTATCTGGTACGACA	ca
1.000	H	3R 5926660 5926759 +	miR-34*			12+	109	TGGCAGTGTGGTTAGCTGGTT	hc
1.000	H	3R 6233855 6233952 +	miR-994			12+	65	CTAAGGAAATAGTAGCCGTGAT	
1.000	H	X 15341899 15341993 +	miR-283	14->15	l(1)G0168	12+	275	AAATATCAGCTGGTAATTCTGGG	
1.000	HM	X 15342890 15342990 +	miR-304*		l(1)G0168	11+	96	TAATCTCAATTTGTAAATGTGAG	hc
1.000	HM	X 15343411 15343482 +	miR-12*		l(1)G0168	11+	129	TGAGTATTACATCAGGTACTGGT	d
1.000	H	X 15799886 15799976 -	miR-927			12+	134	TTTAGAATTCCTACGCTTTACC	
1.000	H	X 17958301 17958382 +	miR-969			12	34	GAGTTCCACTAAGCAAGTTTT	
0.998	HM	2R 12346295 12346381 +	miR-8*			12+	282	TAATACTGTCAGGTAAAGATGTC	hca
0.998	HM	2R 16120930 16121017 +	miR-7*		bl	12+	95	TGGAAGACTAGTGATTTTGTTGTT	ha
0.998	LM	2R 8845317 8845400 -	miR-184*			12+	44	TGGACGGAGAACTGATAAGGGC	ha
0.998	HM	2R 9730653 9730740 -	miR-308*		RpS23	12+	244	AATCACAGGATTATACTGTGAG	dca
0.998	HM	3L 625742 625820 +	bantam*			12+	82	TGAGATCATTTTGAAAGCTGATT	ca
0.998	HM	3R 11243121 11243206 -	miR-13b-1*			12+	294	TATCACAGCCATTTTGACGAGTT	dca
0.998	HM	3R 11243259 11243343 -	miR-13a*			12+	299	TATCACAGCCATTTTGATGAGTT	dca
0.998	HM	3R 21472230 21472315 +	miR-92a*		CG17383	12+	203	CATTGCACTTGTCCCGGCCTAT	dhca
0.996	L	2L 19565419 19565490 -	miR-2a-2*	44->46	spi	12+	64	TCACAGCCAGCTTTGATGAGCTA	d
0.996	LM	2L 19565824 19565903 -	miR-2a-1*		spi	12+	293	TATCACAGCCAGCTTTGATGAGCT	dca
0.994	HM	2L 243037 243141 -	miR-965		kis	12+	73	TAAGCGTATAGCTTTTCCCCTT	
0.994	H	2R 13298658 13298749 -	miR-31a*			12+	69	TGGCAAGATGTCGGCATAGCTG	dc
0.994	HM	3L 22452052 22452119 -	miR-957			12+	45	TGAAACCGTCCAAAACTGAGG	
0.994	HM	X 1645021 1645108 -	miR-981			12+	147	TTCGTTGTCGACGAAACCTGCA	c
0.992	H	2L 7425966 7426051 +	miR-305*			12+	183	ATTGTACTTCATCAGGTGCTCT	ca
0.992	HM	3L 10339177 10339259 +	miR-276a*			12+	86	TAGGAACTTCATACCGTGCTCT	d
0.992	HM	3R 6234025 6234090 +	miR-318*			12+	52	TCACTGGGCTTTGTTTATCTCA	dc
0.990	HM	2L 8258612 8258696 -	miR-2b-1*			12+	293	TATCACAGCCAGCTTTGAGGAGCG	dca
0.990	HM	2L 857543 857629 +	miR-375			12+	196	TTTGTTCGTTTGGCTTAAGTTA	h
0.990	H	X 17962163 17962235 +	miR-210*	46->45;46		12+	81	CTTGTGCGTGTGACAGCGGCTAT	
0.988	H	3L 15808674 15808746 -	miR-263b*		CG32150	11+	130	CTTGGCACTGGGAGAATTCAC	h
0.988	H	3R 12681996 12682065 +	miR-iab-4*			12+	65	ACGTATACTGAATGTATCCTGA	a
0.988	HM	X 12896014 12896100 +	miR-971			11+	121	TTGGTGTTACTTCTTACAGTGA	
0.988	HM	X 8936443 8936532 +	miR-13b-2*		CG7033	12+	294	TATCACAGCCATTTTGACGAGTT	dca
0.986		3L 4989729 4989829 +	Novel-17			10+	519	AAAATTATGCGGAAACGGAAGC	
0.986	H	X 12470276 12470364 +	miR-970		tomosyn	12+	52	TCATAAGACACACGCGGCTAT	
0.982	L	2R 11580121 11580206 -	miR-137			12	256	TATTGCTTGAGAATACACGTAG	h
0.982	HM	3R 17623951 17624051 -	miR-999		Caki	12+	115	TGTTAACTGTAAGACTGTGTCT	
0.980		3L 10936322 10936415 +	Novel-21			12+	26	TCGTCGCATGCGCGTGATCAAC	
0.980	HM	3L 19530386 19530475 +	miR-9a*			12+	194	TCTTTGGTTATCTAGCTGTATGA	dhca
0.980		3R 23797295 23797385 -	Novel-22		betaTub97EF	9	521	TTTATTGCGGCCTGGCCTGACA	
0.978		2R 10136644 10136747 +	Novel-23			12+	105	GAAAGAATAAGAACGGCCAACT	h
0.978	HM	2R 15176452 15176532 -	miR-3*			8+	52	TCACTGGGCAAAGTGTGTCTCA	dc
0.978	LM	3L 11630870 11630963 +	miR-274	14->15	CG32085	11+	89	TTTGTGACCGACACTAACGGGTA	
0.978	H	3L 19731817 19731898 +	miR-33	11->13	HLH106	12+	117	GTGCATTGTAGTCGCATTGTC	h
0.978	H	3R 8377236 8377341 -	miR-284	69->73		12+	65	GTCAGCAACTTGATTCCAGCA	h
0.976		3R 22570390 22570454 +	Novel-24			12+	98	TCAATCAAATCACATGACTGCT	
0.976		3R 24822601 24822689 -	Novel-25		CG1443	12+	292	TGCATTTAAGCCAATTAGCATA	
0.974	H	2L 5642112 5642202 +	miR-964		CG31646	6+	146	TTAGAATAGGGGAGCTTAACTT	
0.974		3L 7188122 7188227 -	Novel-27			5	23	TGAGTCCTTTCACTGGCCACTC	
0.970	HM	2L 16694323 16694425 +	miR-306*		grp	7+	47	TCAGGTACTTAGTGACTCTCA	da
0.968		2L 11749802 11749899 -	Novel-28			8+	109	TGCTTTGAGGTTTATTAGCTGC	
0.968	H	3R 17447616 17447730 -	miR-998		E2f	12+	39	TAGCACCATGAGATTCAGCTC	dhc
0.966	HM	3R 16561652 16561725 +	miR-995		cdc2c	10+	44	TAGCACCACATGATTCGGCTT	dhc
0.966	HM	3R 21477135 21477201 +	miR-92b*			12+	277	AATTGCACTAGTCCCGGCCTGC	dhca
0.964		3R 16281787 16281884 +	Novel-31			12+	68	AATGTCATTAAATTCTCATACA	
0.962		X 15110474 15110559 +	Novel-32			11+	1399	TTTTATTTGTGTCACTGAGTGG	
0.960		3R 21923504 21923596 +	Novel-33			12+	178	TTTGTTCGAGTTGACGTTTGGA	dh
0.960		X 18015331 18015398 -	Novel-34			10	320	TACATAATGTCTCTGTAGGCC	

Score	Val	Locus	Name	MatureChange	Host Gene	Species	Targets	Mature	Orthologs
0.958	HM	2L 18467963 18468040 +	let-7*			12+	63	TGAGGTAGTAGGTTGTATAGT	dhca
0.958		2L 3858079 3858150 +	Novel-35			12+	482	AATTTAATGTGTCGGCGTGTTT	
0.958		2R 13077473 13077566 +	Novel-36		Klp54D	8+	59	TGTTCTCTCCCATTTCTGACTC	
0.956	HM	3R 17448218 17448294 -	miR-11*		E2f	10+	161	CATCACAGTCTGAGTTCTTGCT	dca
0.954	H	2L 13747747 13747840 -	miR-968			6+	85	TAAGTAGTATCCATTAAAGGGTTG	c
0.954	H	3R 9289945 9290028 -	miR-252		CG17025	12+	190	CTAAGTACTAGTGCCGCAGGAG	
0.952		2L 15654271 15654348 -	Novel-39			11+	146	TAATTGCCTGTAAACATAAAGG	c
0.952	HM	2L 6902071 6902141 +	miR-932		neuroligin	8+	90	TCAATTCCGTAGTGCATTGCAG	
0.952		3L 7528550 7528614 -	Novel-41			12+	193	TACTTTTACTTTCATTTATCAA	
0.938	LM	2L 16694680 16694754 +	miR-9b*		grp	11+	253	TCTTTGGTGATTTTAGCTGTAT	dhca
0.932	HM	3L 22603142 22603210 -	miR-958			12+	131	TGAGATTCTTCTATTCTACTTT	
0.924	H	2L 5640944 5641034 +	miR-959		CG31646	9	92	TTGTCATCGGGGGTATTATGAA	
0.922	HM	X 136993 137094 -	miR-980		CG3777	12+	89	TAGCTGCCTTGTGAAGGGCTTA	h
0.900	H	X 19392970 19393056 +	miR-977			5	73	TGAGATATTCACGTTGTCTAA	
0.878	HM	2R 11171939 11172014 +	miR-278*			12+	34	TCGGTGGGACTTTCGTCCGTTT	a
0.864	H	X 19392840 19392924 +	miR-976			5	93	TTGGATTAGTTATCATCAATGC	
0.860	H	X 19392692 19392776 +	miR-975			6		TAAACACTTCCTACATCCTGTAT	
0.856	LM	2L 5641064 5641155 +	miR-960		CG31646	6	127	TGAGTATTCCAGATTGCATAGC	d
0.816	LM	3R 2602061 2602168 +	miR-993			12+	27	GAAGCTCGTCTCTACAGGTATCT	c
0.796	H	2L 18468243 18468354 +	miR-125			12+	24	TCCCTGAGACCCTAACTTGTGA	hca
0.780	H	X 19385468 19385560 +	miR-973			4		TGGTTGGTGGTTGAACTTCGATTTT	
0.752	H	2L 5641993 5642074 +	miR-963		CG31646	8		ACAAGGTAAATATCAGGTTGTTTC	
0.752	HM	2R 16098913 16098990 -	miR-312*			5	295	TATTGCACTTGAGACGGCCTGA	dhca
0.740		2R 3809920 3809997 +	miR-280	9->DEL		11+			
0.738	H	2L 13747605 13747706 -	miR-1002			5		TTAAGTAGTGGATACAAAGGGCGA	c
0.724	H	X 4213551 4213632 -	miR-983-1		CG3626	2		ATAATACGTTTCGAACTAATGA	dh
0.724	H	X 4213730 4213811 -	miR-983-2		CG3626	2		ATAATACGTTTCGAACTAATGA	dh
0.704		2L 20605951 20606060 -	miR-288	71->DEL		12+			
0.704	H	3L 3282582 3282683 +	miR-955			7		CATCGTGCAGAGGTTTGAGTGTC	
0.676	HM	2R 15176569 15176637 -	miR-309*			8	52	GCACTGGGTAAAGTTTGTCCTA	d
0.610	H	X 19395181 19395254 +	miR-978			5		TGTCCAGTGCCGTAAATTGCAG	
0.604	HM	2R 15175726 15175813 -	miR-6-2*			12+	299	TATCACAGTGGCTGTTCTTTTT	dca
0.556	HM	3R 11243462 11243573 -	miR-2c	69->71		12+	63	TCACAGCCAGCTTTGATGGGCA	d
0.554	HM	2R 16098606 16098716 -	miR-310			5	295	TATTGCACACTTCCCGGCCTTT	dhca
0.458	H	X 19385093 19385177 +	miR-972			3		TGTACAATACGAATATTTAGGC	
0.440	L	2L 18467369 18467437 +	miR-100			12+	7	AACCCGTAAATCCGAACTTGTG	hca
0.362	HM	X 8838541 8838640 -	miR-31b*		CG10962;rdgA	11+	68	TGGCAAGATGTCGGAATAGCTGA	dc
0.278	H	3R 23468191 23468275 -	miR-1001		tau	5		TGGGTAAACTCCCAAGGATCA	
0.212	H	X 19395783 19395866 +	miR-979		Grip84	2		TTCTTCCCGAACTCAGGCTAA	
0.198	H	2R 3956799 3956892 +	miR-986		Cyp4e2	5+		TCTCGAATAGCGTTGTGACTGA	
0.166	HM	2R 16098739 16098824 -	miR-311*			5+	295	TATTGCACATTCACCGGCCTGA	dhca
0.132	H	2L 5641201 5641270 +	miR-961		CG31646	4		TTTGATCACCAGTAACTGAGAT	
0.082	H	2R 16100061 16100159 -	miR-991			4		TTAAAGTTGTAGTTTGGAAAGT	
0.070	H	2R 16100169 16100278 -	miR-992			5		AGTACACGTTTCTGGTACTAAG	
0.048		3L 13594772 13594863 +	miR-289	13->DEL	bru-3	12			a
0.042	H	X 4211048 4211134 -	miR-982			4		TCCTGGACAAATATGAAGTAAAT	
0.018	HM	3L 11746959 11747048 +	miR-314*			10+	131	TATTCGAGCCAATAAGTTCGG	
0.012	HM	2R 16099047 16099130 -	miR-313			3	61	TATTGCACTTTTCACAGCCCGA	dhca
0.004	H	2L 12459993 12460110 -	miR-967		bun	5		AGAGATACCTCTGGAGAAGCG	
0.004	L	2L 5641301 5641386 +	miR-962		CG31646	5		ATAAGGTAGAGAAATTGATGCTGTC	
0.000	L	X 4211411 4211490 -	miR-303*	12->14		2	48	TAGGTTTCACAGGAAACTGGTT	
0.000	H	X 4213927 4214027 -	miR-984		CG3626	3		TGAGGTAAATACGGTTGGAATTT	dhca

Supplementary Table S4e: High-scoring miRNA star arms

Name	Mature Count	Star Count	Mature Score	Star Score
miR-9c	1118	104	2.56756	2.51
miR-5	4347	1142	1.98456	3.85226
miR-4	1723	336	2.26861	2.40457
miR-281-2	174	90	0.813321	2.3468
miR-929	112	1	1.7598	3.12731
miR-10	189	1342	1.43093	2.33258
miR-304	69	8	3.02249	2.17256
miR-969	10	0	-0.267922	2.26747
miR-7	1739	96	2.27552	2.22555
miR-9a	4325	95	3.24014	2.44882
miR-959	43	8	2.54689	2.00578

Rfam miRNAs, the cloning frequencies (read counts[40]) and scores during mature prediction (z-scores).

Supplementary Table S4f: High-scoring anti-sense miRNAs

Anti-sense score	miRNA	Validation
1	mir-124	+
1	mir-87	
1	mir-5	
1	mir-307	+
1	mir-1003	
0.998	mir-263a	
0.998	mir-79	
0.998	mir-8	
0.998	mir-184	
0.998	mir-285	
0.998	mir-1041	
0.996	mir-9c	
0.996	mir-2a-2	
0.996	mir-2a-1	
0.994	mir-276b	
0.994	mir-1052	
0.994	mir-92a	
0.992	mir-317	
0.99	mir-275	
0.986	mir-315	
0.984	mir-1027	
0.984	mir-210.1	
0.982	mir-1048	
0.982	mir-iab-4	+
0.978	mir-274	
0.976	mir-13b-2	
0.972	mir-305	+
0.968	mir-2b-2	
0.95	mir-10	

S5 Regulatory motif and instance prediction

S5a Supplementary Methods

Assessing motif conservation by the Branch-Length-Score (BLS) measure

To score evolutionary conservation of short regulatory motifs across many species, we developed a phylogenetic framework that tolerates motif movement and loss, while recognizing their clear selective pressure across the phylogenetic tree[42]. We determined all motif instances in each of the aligned genomes separately and mapped their positions to the alignment. For each motif instance in *D. melanogaster*, we recorded all instances in the other genomes that were aligned, allowing for motif movements in a given window (counted as aligned characters excluding gaps; size of the window depends on the application as described below), but prevented double-counting by assigning each instance in an informant species to the closest instance in *D. melanogaster*. We then evaluated the total evolutionary branch length of the sub-tree of species with conserved motif instances. The overall score of a motif instance becomes this total branch length of the phylogenetic tree over which the motif is conserved, which we call the Branch-Length-Score, or BLS (Figure S5b). The BLS value of a given motif instance ranges from BLS=0.0 (non-conserved) to BLS=1.0 (fully-conserved), representing the fraction of the total phylogenetic tree covered by the species containing the motif.

This BLS conservation measure is applicable to sequence alignments of any set of species and to different types or representations of motifs (e.g. regular expressions, PWMs, or miRNA target sites with and without contribution from 3'end pairing, etc). Importantly, because missing instances in the aligned species are not explicitly penalized, the BLS measure is robust against missing sequence due to low coverage sequencing, assembly errors or alignment artifacts. Lastly, BLS provides a direct estimate of the expected neutral divergence of the species compared[43], accounting for different divergence times between species, correcting for redundant contributions of individual species in a complex tree, and their different rates of divergence (Figure S5b).

Motif-discovery pipeline

We discover motifs with an algorithm that allows us to evaluate the genome-wide conservation of all motifs between 6 and 26 nucleotides with degenerate nucleotides at each position (similar to[6,7,37]; Figure S5b). Due to different background levels of conservation and nucleotide biases, we searched regions of related function separately. This also enables us to discover subtle motifs, which would be overpowered by noise in a more general search. For example, a motif that is only functional in promoter regions may show 30% conservation in promoters, but only 3% conservation across the entire genome, as promoters constitute only 1/10th of all intergenic regions. We first scan the respective regions of the *Drosophila melanogaster* genome for all possible motif-cores (or mini-motif[7,37,44]) of the form ABC-gap-XYZ, where A,B,C,X,Y,Z are each one of the four nucleotides[ACGT] and the gap is of variable length between 0 and 10 nucleotides. For each of the 45056 possible motif-cores, we calculate the total number of occurrences and the number of occurrences with BLS ≥ 50% (90% for transcribed regions) of the 12 *Drosophila* species phylogeny. This cutoff allows for missing data in many different species, e.g. due to sequencing, assembly, or alignment artifacts. We require conserved occurrences to

be aligned in the orthologous regions of a whole-genome alignment, but correct for alignment errors, insertions, and deletions. We measure the genome-wide conservation rate of a motif as the ratio of conserved versus total instances, corrected to the lower bound of a confidence interval (Wilson correction[45,46]). We then determine the *motif excess conservation* (MEC) by subtracting the average conservation rate for random motifs with identical length and nucleotide composition. We also determine the significance of genome-wide conservation by the motif conservation score (MCS) described previously[7,37]. The MCS represents the number of standard deviations by which the observed conservation rate of a motif exceeds the conservation level of random motifs.

We extend the motif-cores into full motifs according to preferential conservation in the gap-region and within 5 nucleotides on both sides of the motif-cores. For this, we evaluated whether specifying additional positions in the gap- or flanking regions with one of the 15 characters from the IUPAC Code improved the discrimination between conserved and non-conserved motif instances. We repeatedly chose and specified the best extension (assessed by a hypergeometric P-value of at least $1x10^{-3}$), until no significant improvement was possible.

We clustered and collapsed motifs with high sequence similarity using hierarchical-clustering with a Pearson correlation coefficient cutoff of 0.77. We calculated a cluster consensus motif by averaging the aligned motifs at each position, which are the final motifs.

Motif-discovery in Promoters, 5'UTRs, Introns, and intergenic regions

We discovered motif in Promoters, 5'UTRs, Introns, and intergenic regions using motif-cores of the form ABC-gap-XYZ as described above (BLS cutoffs 90% for 5'UTRs and CDS and 50% otherwise; MEC cutoffs adjusted based on region sizes to 0.2 (introns, intergenic), 0.1 (enhancer, promoter), 0.04 (5'UTR)). We compared each final motif to known motifs using 0.77 as a cutoff. We assessed the extent of chance matching by comparing the same number of random motifs to the known motifs. The difference between both sets of motifs was assessed by a binomial P-value.

Motif-discovery in 3'UTRs

In 3'UTRs, we used motif-cores of the form ABC-gap-XYZ and continuous 7mers for the discovery pipeline above (BLS cutoff 90%; MEC cutoffs adjusted based on information content to 0.02 and 0.04, respectively). The ABC-gap-XYZ motifs are expanded to full motifs, but are not permitted to have a central gap ≥ 2 after expansion. We assigned matches to miRNA 5'ends when the motifs shared a contiguous 6- or 7mer complementary to positions 1-8 of the mature miRNA. We then tested if motifs constituted novel 5'ends of our predicted miRNAs above, or matched to known non-miRNA motifs. The remaining motifs that were unlikely to constitute miRNAs were clustered based on sequence similarity to remove redundancy.

Prediction of individual motif-instances

We searched all motif instances in the *D. melanogaster* genome and evaluated their conservation with the BLS measure using the 12 *Drosophila* whole-genome alignments (Figure S5b; the phylogenetic tree was based on the whole genome alignment[47]). To evaluate the statistical significance of motif conservation (or the motif *confidence*) in a region-specific way, we created 100 shuffled control

sequences for each TF motif and region and selected those that had a similar number of matches to the region in the *melanogaster* genome (+/- 20%) and were least similar to known motifs. Requiring a similar number of matches to the *melanogaster* genome (i.e. without conservation) controls for di- and tri-nucleotide or other compositional biases as discussed in[48]. To remove possible redundancy among the control motifs, we clustered them (cutoff 0.8), selected only one representative per cluster. For miRNA motifs, we selected all 7mers with identical nucleotide composition and equivalent numbers of matches to the *melanogaster* region (+/- 20%) that were not known miRNA motifs themselves. We then determined total number of instances and conserved instances for all BLS cutoffs (0-100% in steps of 1%) per region for each motif and its shuffles. We calculated a control (random) conservation ratio from the cohort of controls, and used this to determine the expected number of conserved instances (noise). We then defined the confidence for each motif and BLS cutoff (i.e. a motif-specific unique mapping from BLS to motif confidence) as the number of conserved instances above noise divided by the total number of conserved instances. For each motif, region, and confidence, we determined the allowable motif-movement in the alignment (up to 500 nts) that recovered most motif-instances above noise, i.e. do not impose an artificially fixed cutoff.

Motif-representation

We define 'motif' as a string that summarizes the binding specificity of a regulator as inferred from many individual binding sites, or determined experimentally (e.g. by Selex, PBMs). In contrast, a 'motif instance' – or 'regulatory region' – is an individual site of regulator binding, at a fixed position in the genome, and thus is generally much more specific than the motif itself.

We represent all motifs as consensus sequences over an alphabet of 15 characters (IUPAC code, http://www.chem.qmul.ac.uk/iupac/) consisting of the four nucleotides A,C,G,T, the six two-fold degenerate characters S=(CG), W=(AT), Y=(CT), R=(AG), M=(AC), K=(GT), the four three-fold degenerate characters H=(ACT), B=(GCT), V=(G,A,C), D=(G,A,T) and the four-fold degenerate character N=(ACGT). A motif instance or occurrence is a sequence that matches the motif at each position, i.e. contains one of the allowed characters at that position. We obtained TF motifs from Transfac[49], Jaspar[50], FlyReg[51], and the literature. We define miRNA motifs as the non-redundant set of 7mers reverse complementary to Rfam[26] miRNA 5'ends positions 2-8 (seeds,[48]).

Genome annotations

We use the *Drosophila melanogaster* genome-annotations from FlyBase (Release 4.3), and excluded simple repeats and repeatmasker regions obtained from UCSC[5] and non-coding exons according to FlyBase 4.3. We defined promoter regions as the 2000 (100 for core promoters) nucleotides genomic sequence upstream of a transcription start site, truncated to the start or end of the flanking gene. In addition, we searched the genomic regions 5'UTR, protein-coding exon, intron, 3'UTR according to FlyBase 4.3, and treated the remaining sequence as intergenic sequence.

Comparing and Clustering TF motifs

To compare TF motifs, we converted the consensus sequences to position-specific weight matrices (PWMs), concatenated all columns to one vector per motif and calculated the Pearson correlation coefficient between these vectors, filling motif overhangs with Ns to account for variations in motif

lengths and offsets. The similarity score between 2 motifs is the Pearson correlation at the best possible offset between both motifs[52], used and discussed in[37,53,54]. To remove redundancy, we clustered TF motifs using centroid-linkage hierarchical-clustering with a Pearson correlation coefficient cutoff of 0.77. To avoid the creation of artificial motif averages, we selected the one original motif from each cluster that is closest to the cluster-average.

Comparison with experimental datasets

We obtained all experimentally validated miRNA target gene pairs from Tarbase[55] and our previous study[38]. We obtained ChIP-on-Chip regions and the subset that overlapped known enhancers from[56-58] and CrebA target genes from[59]. We calculated the enrichment of sites at different confidence cutoffs between 3'UTRs of validated miRNA/target pairs and all 3'UTRs, and between ChIP regions within 2kb upstream regions and the union of all 2kb upstream regions Fro CrebA and Mef-2, we included the 5'UTR and restricted the upstream region to 500bp instead. We assessed the recovery of motif-instances as the fraction of motif-instances in in 3'UTRs of validated miRNA/target pairs, ChIP regions in 2kb upstream regions that overlapped known enhancers, or 500nts upstream region of validated CrebA target genes, that reached the indicated confidence. To assess the fraction of these that are expected by putatively increased overall conservation in these regions, we assess the recovery of control motifs at the same BLS (not confidence, as the control motifs – by definition – would not reach high confidence levels).

Target genes

We assigned target genes to motif instances when the motif instance overlapped the promoter region or the 5'UTR (TF motifs), or the 3'UTRs (miRNA motifs).

Functional analysis

We obtained functional annotation for *Drosophila* genes from BDGP (in situ expression data; ImaGO), gene ontology (GOslim), KEGG, and FlyBase (genetic interactions, GI). To assess if a motif is related to a functional category, we determined if its conserved (50% BLS) or total instances are over-represented or depleted in promoters and introns of genes from that category compared to those of all genes. We used control motifs to correct for general differences in lengths and conservation between categories that can obscure functional analyses[60]. Over-representation and depletion are assessed by a hypergeometric P-value.

Motif-depletion in CDS and 3'UTRs

We determined the number of instances for each motif and 3 control motifs at 50% BLS in the entire genome and in CDS or 3'UTR regions. We then assessed significant depletion of motif instances in CDS and 3'UTRs by a P-value based on the hypergeometric distribution (cutoff $P=10^{-5}$). We repeated the analysis for the same number of random motifs and assessed the difference between the two motif sets by a hypergeometric P-value.

Motif-multiplicity

We assessed the tendency of motif-instances to cluster by preferential conservation in the context of a second motif-instance. For this, we determined all motif instances in promoter regions that were within 500 nts of a second motif instance, independent of their conservation level. We then assessed if

significantly more such instances were conserved (50% BLS) than isolated instances (i.e. those more than 500 nts from another instance) by a hypergenometric P-value (cutoff $P=10^{-2}$).

Distance bias

To determine if a motif is enriched in regions upstream of transcription starts, we assessed enrichment in such regions compared to the entire genome similar to the functional analysis above. We also assessed if motifs exhibit very defined distance biases at specific positions around the transcription start site, as for example expected for core promoter motifs involved in transcription initiation. For this, we calculated motif enrichment in each of 20 defined 20-nts bins at different positions (-200 nts to +200 nts relative to the TSS) compared to the other bins.

Comparison to phastCons elements

We obtained phastCons elements from UCSC[5] and intersected them with all functional elements from FlyBase and Rfam, and those predicted in our study as described in the respective supplemental results.

Regulatory network properties

Assessing the indegree distribution. We assessed the non-randomness of the indegree distribution against a control Erdos-Renyi random network[61] with the same number of edges network. To construct this network, we added edges by selecting a source and target node with probability 1/m and 1/n, where m and n were the number of source and target nodes in the true network, respectively. We assessed the difference of indegree distributions between the true and control network with a Wilcoxon rank sum test. We also assessed the difference in indegree distribution between all transcription factors (as defined by[62]) and all other genes with a Wilcoxon rank sum test.

Functional/Imago enrichment of high and low indegree genes. We considered all genes with a GO/ImaGO functional annotation (n=7,495 and 5,996, respectively) and computed the indegree (number of incoming edges) for each gene in the transcription factor (TF) and miRNA networks. For both networks we defined high indegree nodes as the 1% with the highest indegree (>=20 for the TF network and >=4 for the miRNA network) and low indegree nodes as miRNA anti-targets (indegree=0) and the same fraction of nodes with lowest indegree in the TF network (80%; <=7 edges). For each GO/ImaGO category, we assessed over-representation and depletion with a binomial P-value.

Mutual Enrichment between high indegree transcriptional and miRNA targets. We considered all genes that were either a target or a regulator in the TF and microRNA networks resulting in a total of 8760 nodes and defined high and low indegree sets as above. We then evaluated if nodes in the miRNA network with high indegree were enriched high indegree nodes of the transcriptional network (or vice versa) using a binomial P-value.

Tissue co-expression. For each TF with available expression information (42; ImaGO), we counted the number of targets that was co-expressed with the TF any of the annotated tissues and the number of targets that was not co-expressed. The statistical significance of co-expression of a TF with its target was estimated using the binomial distribution with p being the probability of a gene being present in one of the tissues in which the TF is known to be expressed, x to be the total number of co-expressed targets and n to be the total number of targets of the TF with known tissue expression.

Network figure. The network figure (Figure S5i) – drawn in Cytoscape[63]- displays genes (nodes) and regulatory connections (edges) of the 60% confidence network. Transcription factors (tfs) with outdegree ≥ are circles, other nodes are squares. Edges are red if the tf and target share at least one tissue according to ImaGO, and nodes are red if they have at least one red edge, otherwise edges and nodes are grey. Bold edges are examples of connections with literature evidence (i.e. is not exhaustive). For clarity, we only show 20 randomly picked targets per transcription factor.

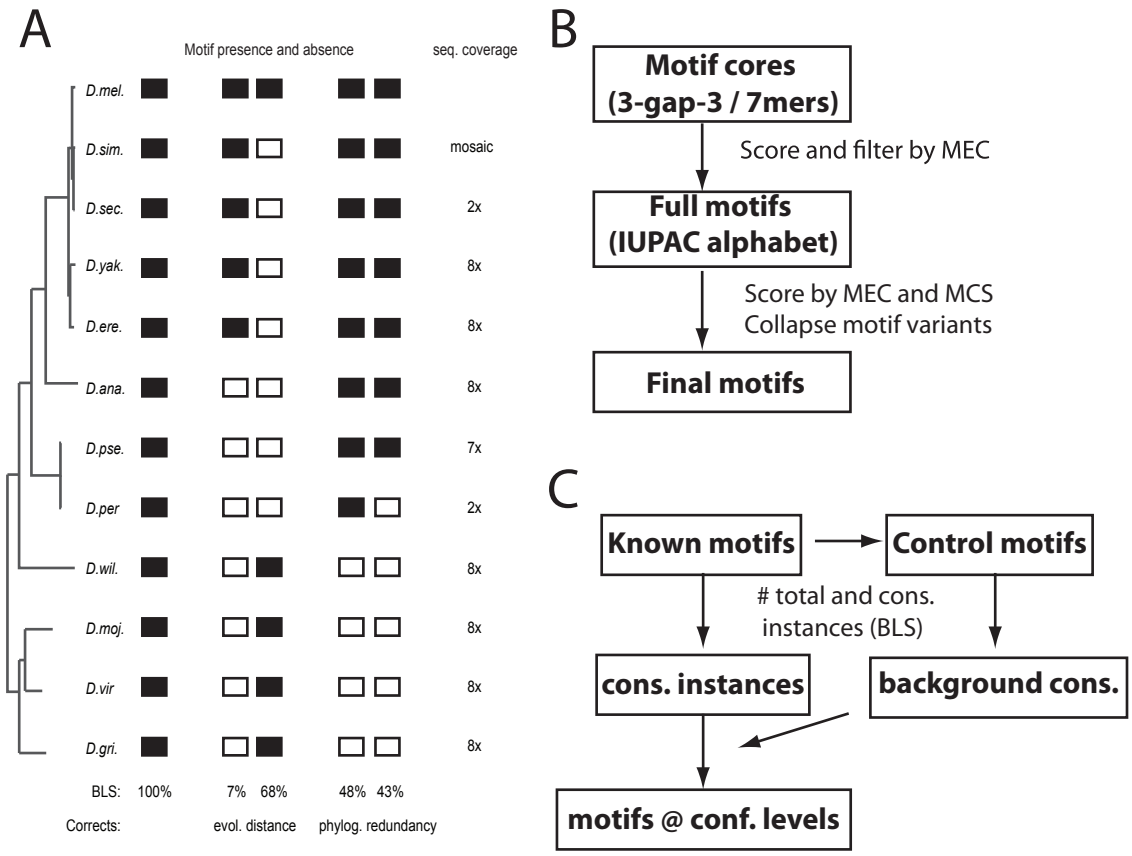


Figure S5b: BLS conservation measure, motif-discovery and motif-instance prediction

(**A**) Calculation of BLS scores for different instance conservation scenarios. Given the pattern of presence (black) and absence (white) within a phylogenetic tree, BLS evaluates the total branch length of the sub-tree connecting the species that contain the motif: when all species are present, BLS is 100%; distantly-related species lead to higher scores as they span larger evolutionary distances ('evol. distance'); species that are very closely related to each other lead to only small incremental contributions, due to their phylogenetic redundancy ('pholog. redundancy'). (**B**) Flow-chart for motif-discovery pipeline. (**C**) Flow-chart for prediction of motif instances.

Supplemental Table S5c. Predicted transcription factor motifs

Name	Motif	MEC	MCS	Origins	Match	MatchName	Multiplicity	ImaGO	ImaGOScore
ME1	GTCACGTD	0.448	45.41	IGP					
ME2	AWNTGGGTCA	0.393	26.97	IPG	GGGTCA	Hormone receptor-like in 46		esophagus (13-16)	4.52
ME3	BCATAAATYA	0.369	36.02	PGIEC	TTTATG	caudal		ubiquitous (13-16)	-6.22
ME4	HAATTAYGCRH	0.365	32.71	PI5GEC	TAATTR	engrailed			
ME5	STATAWAWR	0.358	24.31	C	STATAWAWRSVVV	TATA		ventral nerve cord (13-16)	-5.1
ME6	VATTWGCAT	0.356	44.06	IGPE5			3.73	ubiquitous (11-12)	-7.15
ME7	BYAATTARH	0.338	15.45	GPEIC5	TAATTR	engrailed	7.08	ubiquitous (11-12)	-10.26
ME8	HRTCAATCA	0.338	42.32	PIG				dorsal pharyngeal muscle PR (11-12)	-4.15
ME9	TGACANNNNNNTGACA	0.336	9	G					
ME10	RCGTGNNNNGCAT	0.329	15.94	GPI					
ME11	MATTAAWNATGCR	0.324	12.43	GIP	TGCATAATTAATTAC	abnormal chemosensory jump 6		tracheal PR (11-12)	4.11
ME12	TTAATGATG	0.32	20.31	GP					
ME13	WTGACANBT	0.318	63.45	GPIE5			4.14	ubiquitous (13-16)	-3.97
ME14	YGACMTTGA	0.313	27.06	IPG				midgut (13-16)	4.32
ME15	AATTRNNNNCAATT	0.309	21.17	GP					
ME16	TGACGTCAT	0.304	12.24	5CIGP	TGACGTCA	CrebA			
ME17	MAATTNAATT	0.304	51.57	IGPE5				ubiquitous (11-12)	-6.66
ME18	MRYTTCCGYY	0.304	39.04	PIGE	SGGAAA	dorsal		ubiquitous (11-12)	-4.4
ME19	MATTRRCACNY	0.303	25.24	GPI					
ME20	YTAATGAVS	0.298	44.5	GPEI	VNRYTAATGRBM			foregut PR (11-12)	4.19
ME21	TAATTRANNTTNATG	0.294	8.67	G					
ME22	WAATGCGCNT	0.291	18.17	G					
ME23	MATTWRTCA	0.288	46.25	PGEI				dorsal epidermis PR (11-12)	4.4
ME24	YAATTWNRYGC	0.287	30.91	GP			4.27	ubiquitous (11-12)	-4.79
ME25	TTAYGTAA	0.283	13.06	5	TWTKACKTAANA	giant		midgut (13-16)	5.32
ME26	YGCGTHAATTR	0.283	13.61	GPE					
ME27	AATTRYGWCA	0.28	22.85	IGEP				pericardial cell (13-16)	4.1
ME28	GCGCATGH	0.28	30.17	PCEG				ventral nerve cord PR (11-12)	5.75
ME29	WAATCARCGC	0.275	13.82	G					
ME30	AATTAANNNNNCATNA	0.271	16.44	G	TTAATT	Antennapedia			
ME31	GCGTSAAA	0.271	29.95	GP					
ME32	YGCGYRTCAWT	0.269	12.87	G					
ME33	GCGTTGAYA	0.269	15.1	GP					
ME34	AAATKKCATTA	0.266	14.04	GP					
ME35	RACASCTGY	0.266	28.38	ECPG	GCAGSTGK	scute		ventral sensory complex SA (11-12)	4.08
ME36	TGTCAATTG	0.265	12.65	GP				tracheal system (13-16)	4.56
ME37	WAATKNNNNNCRCGY	0.261	23.34	GEP					
ME38	CASGTAR	0.261	9.24	EPG	GTACGTG	single-minded	4.58	ventral epidermis PR (11-12)	7.41
ME39	WCACGTGC	0.26	10.54	5EPGIC	TGGCACGTGYYA	Enhancer of split			
ME40	CATTANNNWAATT	0.259	19.02	G					
ME41	TYRACACTTK	0.259	21.05	G				adult salivary PR (13-16)	4.14
ME42	RCGCMATTW	0.258	31.06	GP	GCCATT	pleiohomeotic		central brain surface glia (13-16)	4.39
ME43	YAATKAAGY	0.256	36.54	GP	VVDVYAATTAAG			ubiquitous (11-12)	-5.57
ME44	TGACANNTTGAC	0.25	8.42	GP					
ME45	VCACGCRH	0.25	75.8	IEPCG	CACGCG		3.93	ventral nerve cord (13-16)	4.55
ME46	RAGTGAAAGT	0.249	13.03	PI5					
ME47	AATTANWNRCGC	0.248	14.12	G					
ME48	AATTWWNAYGCR	0.245	17.62	GPEI					
ME49	CGTGNGAA	0.244	6.37	GP	BYGTGRGAAMCBNDVD		4.39	ventral epidermis PR (11-12)	4.62
ME50	TRACRYGCA	0.241	28.33	G					
ME51	RCAAWTTR	0.239	84.61	GEPC			6.44	ubiquitous (11-12)	-6.83
ME52	RKGTCAAGK	0.239	28.07	PIGC					
ME53	YGTCWAATTA	0.235	14.99	G					
ME54	RCGYRCGY	0.228	52.73	P					
ME55	TAATATGCRA	0.227	13.34	G					
ME56	TNATGNNTGACA	0.225	12.87	G					
ME57	RCGTGYYAAAT	0.222	9.53	G					
ME58	TAATTNWMATT	0.219	28.37	GP					
ME59	GCGYNWNAWTGAY	0.218	10.86	G					
ME60	CATNANTYAAA	0.217	25.4	G					
ME61	AACWAATTR	0.216	38.85	PG					
ME62	TGGCGCC	0.212	32.57	EPG	TGGCGYY	brinker		dorsal pharyngeal muscle PR (11-12)	4.51
ME63	AAATCAAT	0.21	12.52	GPC5E					
ME64	CNNNGCGYRTGANYNAT	0.209	6.19	GIP				ventral nerve cord PR (11-12)	-4.21
ME65	ACGYNGCGTATGM	0.208	5.7	GIP				ventral epidermis (13-16)	-4.62
ME66	AAAATGCA	0.203	32.17	P5				ubiquitous (11-12)	-4.88
ME67	CATTANYGTCA	0.203	8.56	PG					
ME68	GCATAHWWWNNNGCGY	0.201	7.66	G					
ME69	AATKACA	0.201	61.65	EGP				ubiquitous (13-16)	-4.23
ME70	GACAATK	0.2	47.33	P			3.18	ubiquitous (11-12)	-7.26
ME71	TGTCAMNTGCA	0.194	6.37	G					
ME72	KCAATAAA	0.194	40.64	GPE			2.9		
ME73	AAAGTGANA	0.192	28.95	P			2.92	tracheal system (13-16)	8.51
ME74	YGATAAGC	0.191	21.43	PCG	ANHDDBHGATAASSDNNB	pannier		midgut (13-16)	7.36
ME75	TAATKRNGTCATTA	0.191	5.33	G					
ME76	TGACAWWTWTGC	0.186	6.47	G					
ME77	TGAYWWWTGCA	0.186	13.96	G					
ME78	AATTNNNTCACGY	0.186	8.93	G					
ME79	GCTNMTTAA	0.185	28.52	P			3.09		
ME80	GTCANTNAAC	0.181	17.54	P	VRGKTYAWTGAMMYY	Ecdysone receptor			
ME81	CTCRTAAAW	0.18	19.93	G	TTTATG	caudal		ubiquitous (11-12)	-6.61
ME82	TNAGCATAA	0.174	22.74	P					
ME83	MACMDGTTK	0.171	41.99	5PCEIG				ventral epidermis (13-16)	8.14
ME84	TRTCAWNWRTCA	0.167	13.5	G					
ME85	TAATTRNNNNYGACA	0.167	8.64	G					
ME86	YGTCAWTGAC	0.164	9.16	PG5					
ME87	MACTTGTYR	0.164	20.93	GP	RCARGT				
ME88	ACGNNAATTG	0.16	17.3	P					
ME89	CACRCAC	0.158	59.4	PCE			6.38	ventral nerve cord (13-16)	8.13
ME90	ACATGTK	0.155	38.96	P	CACATGT	twist		ventral epidermis PR (11-12)	4.65
ME91	CATNNNNCGCG	0.155	5.06	E					
ME92	TAACCTC	0.153	7.55	5				ubiquitous (11-12)	6.28
ME93	GCAACA	0.151	60.64	PC	CAACAA	Adult enhancer factor 1	7.87	procephalon PR (11-12)	9.14
ME94	RAGTKCAANG	0.147	17.12	IP					
ME95	AAASTTT	0.146	56.89	PG			10.6	central brain neuron (13-16)	6.59
ME96	RTAAACA	0.144	52.21	P	RTAAACAA			muscle system (13-16)	5.22
ME97	GCGNNTNNTTA	0.143	27.1	E					
ME98	AATTRNNNNNCAA	0.142	31.51	P				ubiquitous (11-12)	-4.92
ME99	GNMCTTGAA	0.135	20.5	P					
ME100	GTATNWATA	0.13	8.82	C	RTATATRTA			central brain PR P1 (11-12)	-3.92
ME101	ATGANNNNTCA	0.128	5.08	E					
ME102	TWAWKMNAWTTG	0.122	19.64	G					
ME103	YRAAMGTGM	0.112	24.42	PG				tracheal system (13-16)	5.85
ME104	CTTNNATAC	0.11	8.12	C					
ME105	GYATGMGWAATKA	0.107	4.81	IG					

Supplemental Table S5c. Predicted transcription factor motifs (cont.)

ME106	GGTNTAAAW	0.106	6.6	C					
ME107	ACTNACCT	0.1	19.82	I5					
ME108	AATNNNNCATNR	0.099	33.83	E					
ME109	ATCWATG	0.099	28.75	5					
ME110	ATCATAA	0.098	23.79	E	TTATS	mitochondrial transcription factor A			
ME111	ATCTNATC	0.097	7.48	C				midgut (13-16)	8.84
ME112	MAACAA	0.094	74.28	PC	CAACAA	Adult enhancer factor 1	5.13	procephalon PR (11-12)	5.42
ME113	ATTNWTTA	0.093	44.54	E	TAATTAA			dorsal pharyngeal muscle PR (11-12)	-4.57
ME114	TTTNGGCGS	0.088	17.33	5					
ME115	GATNTCAT	0.085	20.11	E					
ME116	TGGATTA	0.083	8.19	5	VVVBTAATCC	bicoid		procephalon (13-16)	-4.76
ME117	MAAMNNCAA	0.08	56.88	PC			7.72	procephalon (13-16)	7.9
ME118	GTCNNNNGACA	0.077	4.51	E					
ME119	GNCTANWWATA	0.077	5.34	C	YTAWWWWTAR	Myocyte enhancing factor 2		muscle system (13-16)	4.83
ME120	AACTGA	0.074	10.24	C	HSWAACHGH	ovo		ubiquitous (11-12)	-4.39
ME121	ACANACA	0.073	17.62	PC			7.14	ventral nerve cord (13-16)	8.49
ME122	ACACNNNNRCAC	0.069	17.97	P				ventral nerve cord (13-16)	6.57
ME123	CACNNNNNNACA	0.068	34.33	P			7.39	ventral nerve cord (13-16)	6.66
ME124	GCANGTCC	0.068	5.19	C			2.94	ubiquitous (11-12)	-4.1
ME125	TAAATAG	0.063	6.88	C	RTAAATA	biniou		ventral nerve cord (13-16)	-4.31
ME126	CAANNNAAA	0.061	46.76	P			4.24	ubiquitous (11-12)	-6.46
ME127	CACNNRNNNNNNCAC	0.061	25.49	P			2.96	labial sensory complex (13-16)	5.77
ME128	CACGAGNC	0.06	13.21	C	CACGCGMC	hairy		fat body/gonad PR (11-12)	4.01
ME129	ATGTGAT	0.059	16.03	5				sensory nervous system PR (13-16)	-4.08
ME130	WTGGNNNNNTAAY	0.057	15.97	E					
ME131	ACACNNACAC	0.053	14.31	PC			2.74	tracheal system (13-16)	-6.18
ME132	RCACNNNNNNNCACA	0.053	15.66	P			2.65	tracheal system (13-16)	-4.84
ME133	AAAAGCT	0.053	18.76	C			3.11	midgut (13-16)	4.21
ME134	CAGNNGCA	0.051	24.24	C			5.11	procephalon PR (11-12)	8.16
ME135	AAANNNNNNCAA	0.048	36.32	P			4.29	ubiquitous (11-12)	-5.71
ME136	AAANNNNNNNNNAAT	0.038	32.91	P				ubiquitous (13-16)	-5.77
ME137	GAGAGAG	0.037	18.1	C			4.3	ventral nerve cord (13-16)	7.13
ME138	TCCTNNNNNNNGGA	0.027	4.01	C				dorsal epidermis (13-16)	-4.48
ME139	ATANANNCGC	0.025	7.85	C					
ME140	TCANNNTGGA	0.025	4.64	5					
ME141	AATNNNNNNNAAAA	0.022	14.26	P				ubiquitous (11-12)	-4.18
ME142	CTATNNNAAG	0.016	4.18	C					
ME143	TNRCGCNYNATTAATY	0.011	3.67	G					
ME144	GCANTTGYNWYAATT	0	2.11	G					
ME145	RTTRCGYATRCGCM	0	2.34	PEGI				ubiquitous (11-12)	-5.83

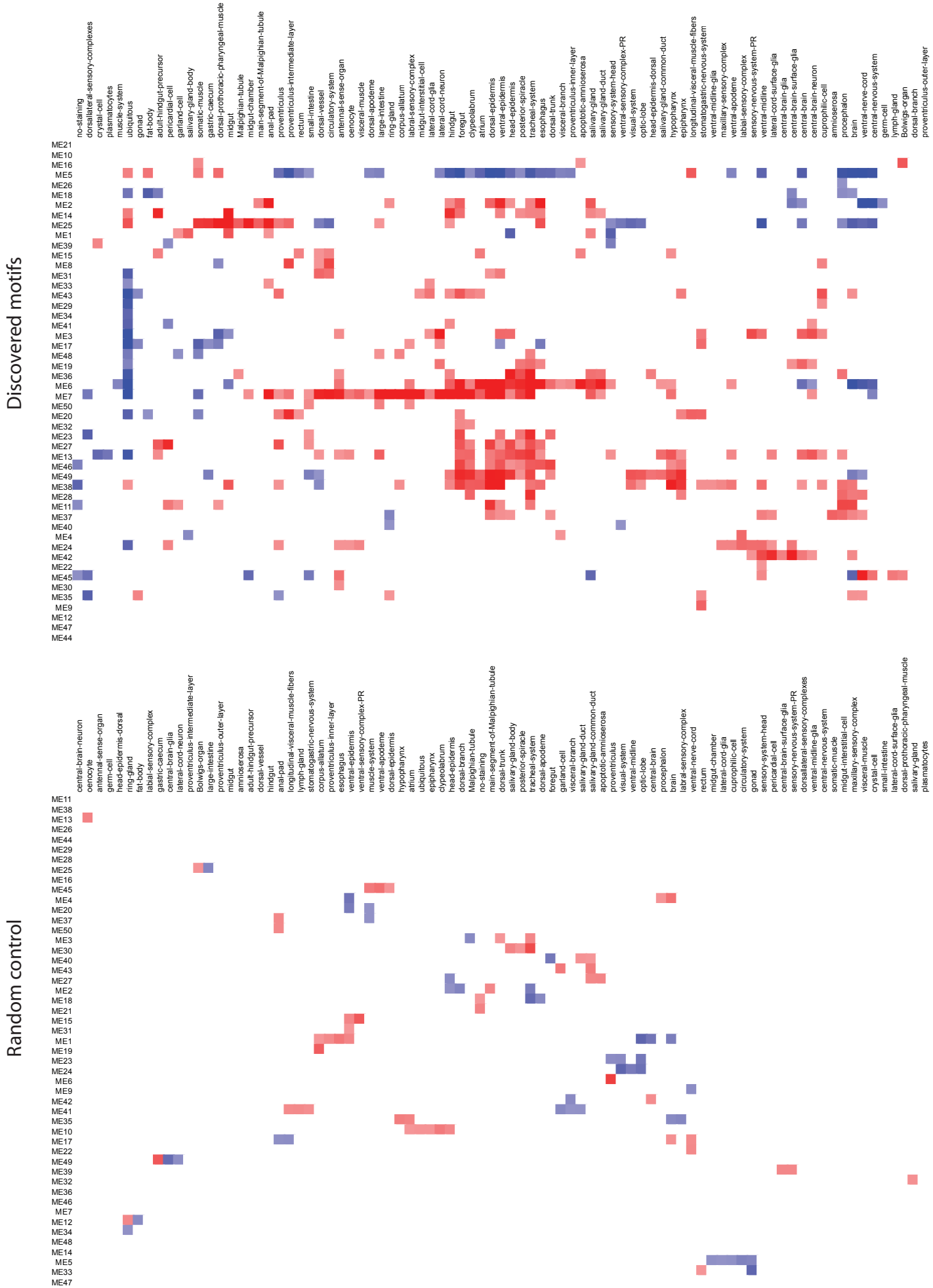
Supplemental Table S5d Recovery of known transcription factor motifs

Motif	MEC	MCS	Length	Factor	Match
TGACGTCA	0.423	26.48	8	CrebA	TGACGTCAT
CAATTA	0.322	14.77	6	reversed polarity	BYAATTARH
CACGCGMC	0.31	29.36	8	hairy	VCACGCRH
TTTATG	0.255	14.05	6	caudal	BCATAAATYA
TTAATT	0.252	16.81	6	Antennapedia	BYAATTARH
TAATTR	0.247	17.64	6	engrailed	BYAATTARH
TYAAGTGS	0.224	46.74	8	ventral nervous system defective	
KYTAATKDNY	0.207	88.72	10	fushi tarazu	BYAATTARH
CAGSTG	0.206	10.97	6	asense	RACASCTGY
GTACGTG	0.204	25.81	7	single-minded	CASGTAR
RRCAGGTGB	0.2	28.7	9	escargot	RACASCTGY
CACTTRA	0.197	52.18	7	tinman	
ATGCGGGY	0.196	21.62	8	glial cells missing	
CAGGTG	0.195	57.85	6	snail	RACASCTGY
AAATTAA	0.193	72.23	7	tailless	BYAATTARH
TCAWTTAAMT	0.188	20.18	10	even skipped	
STATAWAWRSVVV	0.183	7.27	13	TATA	STATAWAWR
TAAT	0.182	31.21	4	apterous	BYAATTARH
VRGKTYAWTGAMMYY	0.175	7.29	15	Ecdysone receptor	GTCANTNAAC
GCAGSTGK	0.173	6.23	8	scute	RACASCTGY
RATTAAW	0.173	97.53	7	retained	BYAATTARH
CACATGT	0.169	26.45	7	twist	ACATGTK
TGGCGYY	0.154	7.9	7	brinker	TGGCGCC
TTATS	0.152	16.84	5	mitochondrial transcription factor A	ATCATAA
GGGTCA	0.15	44.12	6	Hormone receptor-like in 46	AWNTGGGTCA
RWWNTNRCACYT	0.147	24.34	12	brachyenteron	
GCCATT	0.142	56.54	6	pleiohomeotic	RCGCMATTW
CAACAA	0.138	74.55	6	Adult enhancer factor 1	MAACAA
YGYGGTY	0.135	53.63	7	runt	
AWCAGGTGK	0.134	13.23	9	atonal	RACASCTGY
VSNKTDATKRCNV	0.128	32.18	13	Abdominal B	
CMGGAAR	0.123	7.9	7	Ecdysone-induced protein 74EF	
RATTAMY	0.121	64.92	7	Deformed	BYAATTARH
KHGATAASR	0.117	28.43	9	serpent	
TGANTCA	0.106	43.28	7	activating-protein 1 (FOS-JUN heterodimer)	
RTAAATA	0.106	49.54	7	biniou	TAAATAG
WCATTWMM	0.098	46.89	8	zerknullt	YTAATGAVS
KNVNVBYTAATKRSBHNVD	0.098	13.81	19	Ultrabithorax	BYAATTARH
RAAMGGRTTA	0.095	10.25	10	Kruppel	
YTAWWWWTAR	0.081	11.79	10	Myocyte enhancing factor 2	GNCTANWWATA
RTAAMA	0.079	10.06	6	crocodile	
WCYGGTTT	0.078	21.89	8	grainy head	
RRNNNMCACCTGC	0.077	8.31	13	achaete	RACASCTGY
VVVBTAATCC	0.073	15.35	10	bicoid	TGGATTA
AANTNTAATGACA	0.067	4.88	13	empty spiracles	
TTNNRCAATM	0.053	16.42	10	slow border cells	
MVHTAAKCCS	0.045	12.56	10	ocelliless	
BNWDTYGAGTGRNHDD	0.043	7.85	16	zeste	
BYRHBACAAWGTDDB	0.04	6.12	15	doublesex	
HSWAACHGH	0.039	23.28	9	ovo	AACTGA
GRGGTCAYS	0.032	7.27	9	ultraspiracle	
SVTAATYGATTANS	0.031	3.96	14	paired	
CWYBDCY	0.03	51.79	7	prospero	
RKAAASA	0.027	9.05	7	broad	RTAAACA
SGGAAA	0.021	11.77	6	dorsal	MRYTTCCGYY
WHWWWWWWWWK	0.019	29.14	12	bric a brac 1	
SMATAAAAAA	0.017	6.54	10	hunchback	
YYWVNYYWDNYS	0.014	19.95	12	Dorsal interacting protein 3	
TWTKACKTAANA	0.012	4.27	12	giant	TTAYGTAA
GAGGAAGC	0.012	5.53	8	pointed	
RBYGTGRGAAMCB	0.01	4.03	13	Suppressor of Hairless	CGTGNGAA
AAHKMTHBCA	0.003	4.64	10	knirps	
RSWGAGMRHRR	0.001	4.29	11	Trithorax-like	
ANHDDBHGATAASSDNNB	0	3.15	18	pannier	YGATAAGC
BMGYBGYYGYNGMVBV	0	0.29	16	Adh transcription factor 1	
CAATGCACTTCTGGGGCTTCCAC	0	0	23	glass	
CCTTTGATCTT	0	2.3	11	pangolin	
CHGGAW	0	2.48	6	Ets at ???	
CKCAKCWCT	0	2.35	9	Serendipity _	
GGGGAMWWCCM	0	0.12	11	schnurri	
GGGGAWTCCCY	0	0.06	11	disordered facets	

Supplemental Table S5d. Recovery of known transcription factor motifs (cont.)

GGGGAWYCMC	0	0.29	10	Relish	
KVRKRNTCACTSRNTVHDB	0	0.59	19	eyegone	
MGAADMGAADMGAAD	0	2.85	15	Heat shock factor	
MKSCMAGGACVHH	0	2.41	13	tramtrack	
RRAYATTYBKSGVATKVCA	0	0	19	scalloped	
RTATATRTRB	0	0.07	10	Chorion factor 2	STATAWAWR
RWWWASWBDYSKNMW	0	0.25	15	mirror	
TATCGATA	0	2.35	8	DNA replication-related element factor	
TGCATAATTAATTAC	0	0	15	abnormal chemosensory jump 6	MATTAAWNATGCR
TGCTCAATGAA	0	2.84	11	Tor Responsive Element	
TGGAGGDGGWAHTMATBVRTGWDDDRKKMW	0	0	30	twin of eyeless	
TGGCACGTGYYA	0	0	12	Enhancer of split	WCACGTGC
TTCCSGGAA	0	3.23	9	Stat92E	
VSGYYGCMGYCGYYGMMKKYG	0	0	21	Adh transcription factor 2	
YGATAC	0	0	6	sine oculis	
YSAAGGWCRCHRM	0	1.88	13	ftz transcription factor 1	

Supplemental Figure S5e. Tissue enrichment and depletion for discovered TF motifs



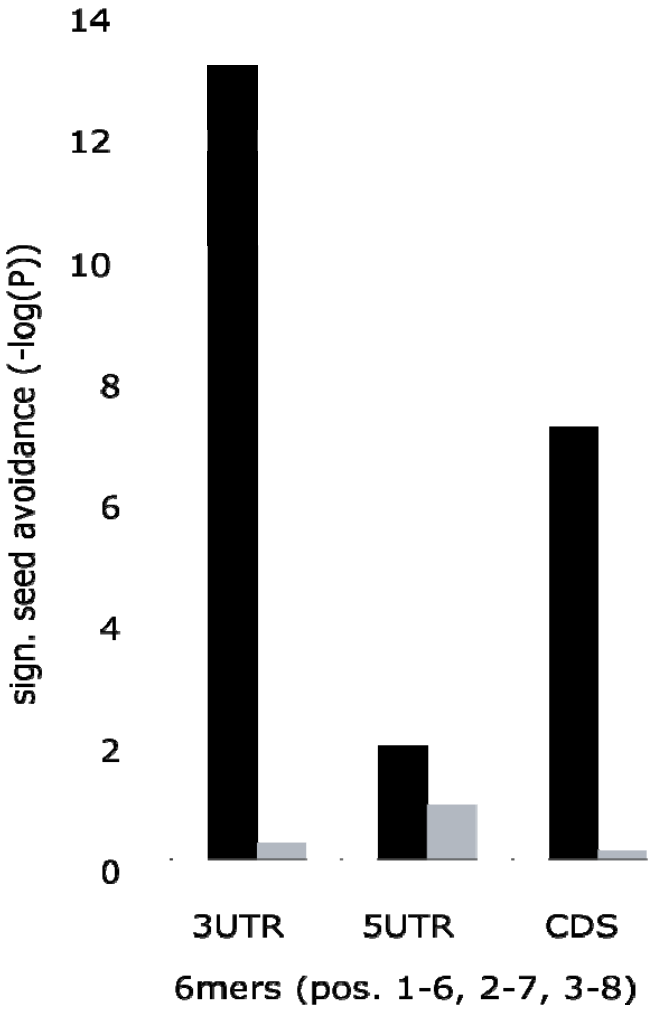


Figure S5f. Depletion of miRNA motifs in 3'UTR motifs of ubiquitously expressed housekeeping genes or genes co-expressed with the miRNA results from such 'anti-target' genes avoiding miRNA-mediated repression[38,64,65]. Depletion of miRNA 6mer motifs (complementary to miRNA 5'ends) in the coding regions of anti-target genes thus suggest that such sites can function biochemically. Shown is the negative logarithm of a hypergeometric P-value indicating significant depletion in anti-target genes (according to[38]) with respect to average genes in 3'UTR, 5'UTR, and coding regions.

Supplemental Table S5g. Predicted 3'UTR and coding region motifs

Name	Motif	Origins	3UTRMCS	3UTRMEC	CDSMEC	CDSfMEC	Known Match
MO1	CTGTGAT	3C	12.86	0.459	0.092	0	miR-2a;miR-2b;miR-6;miR-11;miR-13a;miR-13b;miR-308;miR-2c
MO2	AAGACTG	3C	11.08	0.36	-0.019	0.05	miR-14
MO3	GTGCCAA	3C	10.2	0.34	0.043	0.097	miR-263b
MO4	TCTAGTC	3	10.46	0.322	0	0	miR-279;miR-286
MO5	GTGCCTT	3	9.36	0.32	0	0.059	miR-124
MO6	GTACAAA	3C	14.95	0.297	0.045	0.082	miR-305
MO7	GACAATA	3	12.32	0.297	-0.009	0.033	miR-1000+
MO8	TACCTCA	3	9.22	0.293	-0.119	0.068	let-7
MO9	GTCTTCC	3	8.56	0.284	0.04	0.044	miR-7
MO10	GTGCAAT	3	11.44	0.272	-0.088	0.082	miR-92a;miR-92b;miR-310;miR-311;miR-312;miR-313
MO11	ACATTCC	3C	11.13	0.266	0.108	0.258	miR-1
MO12	TGCATTT	3	14.89	0.26	0.049	0.129	miR-277
MO13	TCCGTCC	3	6.83	0.259	0.021	0.037	miR-184
MO14	GTAAATA	3	18.63	0.258	0	0	miR-280
MO15	GTACCTG	3	5.97	0.257	0.032	0.047	miR-275;miR-306
MO16	CAGTATT	3C	11.1	0.247	0.06	0.041	miR-8
MO17	ACTGCCA	3C	8.85	0.244	0.002	0.046	miR-34
MO18	TAAGTAG	3	11.11	0.231	-0.011	0	
MO19	GTGCCAT	3	6.91	0.23	0.187	0.16	miR-263a++
MO20	ACATATC	3C	10.31	0.226	0.13	0.124	miR-190+
MO21	AGCTTTA	3	11.75	0.226	-0.041	0	miR-4
MO22	ATGTNCCT	3	8.91	0.224	0.059	0.052	
MO23	GTGCNATT	3	8.16	0.22	-0.032	0.073	
MO24	TAATTTAT	3C	11.77	0.216	0.075	0	
MO25	GTGNCTTA	3	7.27	0.212	0	0	
MO26	ACCAAAG	3	11.74	0.208	0.06	0.06	miR-9a;miR-9c;miR-9b
MO27	GTTCCTT	3	7.87	0.208	-0.109	0.001	miR-5
MO28	GTATTAT	3C	13.33	0.205	0.02	0	miR-983-1+;miR-983-2+
MO29	TGCTCAA	3	7.77	0.204	0.13	0	miR-87
MO30	GTACTTA	3	9.95	0.198	-0.055	0	miR-252+
MO31	AAAAGAC	3	12.34	0.19	0.018	0.035	miR-316
MO32	AATAAA	3	37.6	0.19	0	0	
MO33	ATGTGCC	3	6.75	0.187	0.309	0.02	X:22051016-22051080:-
MO34	AATACTC	3	8.84	0.181	0.022	0	miR-12
MO35	GTGGCCTT	3	3.29	0.18	0.115	0.043	
MO36	AAGTNCCT	3	6.99	0.176	0.108	0	
MO37	ATTTATT	3C	16.65	0.175	0	0.008	miR-7*
MO38	CGAATTT	3C	10.07	0.175	-0.119	0	miR-10-3p++
MO39	GTCAATT	3C	7.76	0.166	0	0.032	
MO40	TGTANWTW	3	23.82	0.165	0.059	0.004	
MO41	TTGTGCC	3	5.75	0.164	0.173	0.04	
MO42	TGATCTC	3C	5.22	0.163	0.13	0	bantam
MO43	TGTRWNATA	3	10.85	0.161	0.043	0.001	
MO44	TCTTGCC	3	5.85	0.157	-0.01	0	miR-31b;miR-31a
MO45	GCCAWA	3	18.85	0.157	0.014	0.049	
MO46	GTTGCCT	3	6.17	0.157	0.003	0	
MO47	GTGTCTT	3	5.94	0.156	-0.121	-0.001	X:21874753-21874825:-
MO48	GTGNYAAW	3	13.06	0.154	-0.175	0.035	
MO49	TGTGTTC	3	7.55	0.153	0.08	0.039	miR-317
MO50	TTGTTGC	3	8.43	0.153	0.019	0	3R:1827841-1827943:-
MO51	GCCAAAT	3C	8.95	0.152	0.032	0.086	3L:9127192-9127258:+;2R:10136644-10136747:+
MO52	TATTTAT	3	14.8	0.151	0.023	0.049	miR-283;miR-289
MO53	TAGTGCA	3	5.83	0.135	-0.09	0	
MO54	TGTACTT	3	8.21	0.134	0.103	-0.003	miR-992+
MO55	GCTCAGG	3	4.08	0.133	-0.049	0	miR-125
MO56	GCTGTGA	3C	4.77	0.129	-0.035	0	miR-2a-2++;miR-2c++
MO57	TCTTGTCA	3	3.68	0.128	-0.004	0	
MO58	WTTGTTR	3	15.66	0.128	-0.021	0.005	
MO59	AAGTGCA	3	6.83	0.127	0.149	0	3L:6852152-6852264:+
MO60	TTGNGATT	3	7.04	0.125	-0.117	-0.018	
MO61	GTACTGA	3	4.35	0.124	-0.259	0	
MO62	AGTTCCT	3	4.99	0.123	0.149	0.069	miR-276a;miR-276b
MO63	AGCTNTAA	3	7.58	0.121	-0.104	0	
MO64	CCAGTGA	3	4.93	0.12	-0.056	0.483	miR-3;miR-318
MO65	RAGTCTGW	3	5.37	0.119	-0.132	0	
MO66	TGTATTA	3C	8.49	0.117	0	0	miR-33*
MO67	GNATTTAG	3C	6.55	0.11	0.033	0	
MO68	GGCACGTG	3	2.56	0.109	0	0	
MO69	GGTTGTG	3	4.3	0.107	-0.092	0	miR-307
MO70	TGTACAG	3	4.75	0.107	0.058	-0.003	miR-972+
MO71	TTGTTAA	3	8.48	0.107	-0.056	0	3L:4989729-4989829:+
MO72	GTCACAA	3	4.99	0.104	0.028	0.023	miR-274++
MO73	TGATATT	3	7.24	0.102	0.015	0	miR-283++
MO74	AGACTTG	3	4.41	0.098	-0.26	0	3L:6987670-6987774:-
MO75	GGGACCA	3	3.86	0.096	0	0	miR-133

Supplemental Table S5g. Predicted 3'UTR and coding region motifs (cont.)

MO76	ACTNCCTNC	3	4.68	0.096	0	0	
MO77	AGGTRAGT	C	2.55	0.095	0.33	0	
MO78	GCCAAAG	3C	5.3	0.093	-0.162	0.018	miR-79*
MO79	TGCACAA	3	5.08	0.09	0.275	0	miR-210
MO80	ATGNCAA	C	9.76	0.084	0.077	0.07	
MO81	TWAGTT	3	14.02	0.084	-0.076	-0.058	
MO82	GATGTGA	3	4.6	0.084	0.101	0	
MO83	TTGCTGA	3	4.36	0.082	-0.049	0	miR-284++
MO84	TGTNATAA	C	6.11	0.076	0.105	0	
MO85	AAGTGAT	3	4.79	0.074	-0.132	0	3L:9012036-9012098:-
MO86	GACAATC	3	3.64	0.07	0.093	0.02	miR-219
MO87	TAATTW	3C	12.54	0.06	0	0	
MO88	GTGTCAT	3	3	0.055	0.044	0	
MO89	GCACTGA	3	3.27	0.054	-0.003	0	2L:7339686-7339797:-
MO90	GTACAGT	3	2.88	0.045	-0.112	0	
MO91	AGCGCAAT	3	2.25	0.041	0.018	0	
MO92	ATGNCAT	C	4.72	0.041	0.093	0.17	
MO93	ACTGAGC	3	2.17	0.019	-0.282	0	
MO94	TTATTTT	3	3.22	0.016	0	0.008	
MO95	WCAAGAC	C	2.03	0.009	0.152	0.023	
MO96	TGATGTAY	C	1.76	0.005	0.256	0	
MO97	CAGGTGNG	C	1.6	0.004	0.138	0	
MO98	TACAWNATG	C	0.12	0	0.376	0.022	
MO99	TGANATG	C	0.06	-0.002	0.183	0	

Supplemental Table S5h. Predicted regulatory interactions with literature evidence

TF	TF_CG	Target	Target_CG	Evidence type	First Author, Year	PMID
tin	CG7895	Doc3	CG5093	Same Pathway	Reim & Frasch, 2005	16221729
tin	CG7895	bap	CG7902	Direct	Yin & Frasch,1998	9621427
Kr	CG3340	fkh	CG10002	Same Tissue & Pathway	Michael & Jackle, 1998	1170904
Kr	CG3340	hb	CG9786	Same Tissue	Treisman &Desplan, 1989	2797150
ac	CG3796	emc	CG1007	Same Tissue	Martinez et al., 1993	8497266
ac	CG3796	Chn	CG11798	Direct	Escudero et al., 2005	15703278
					Lammel et al., 2000;	10727857
brk	CG9653	fkh	CG10002	Same Tissue	Abrams EW & Andrew, 2005	15901661
brk	CG9653	Doc3	CG5093	Same Tissue	Hatton-Ellis et al.,2007	17190812
brk	CG9653	Doc1	CG5133	Same Tissue	Hatton-Ellis et al.,2007	17190812
brk	CG9653	htl	CG7223	Gastrulation	Stathopoulos & Levine, 2004	15380237
brk	CG9653	prd	CG6716	Same Pathway	Marty et al., 2000	11025666
sn	CG32858	pyd	CG31349	Same Context	Norga et al., 2003	12932322
CrebA	CG7450	spi	CG10334	Same Pathway	Andrew et al., 1997	9006079
CrebA	CG7450	toe	CG10704	Same Context	Kumar et al., 2004	15514061
Su(H)	CG3497	th	CG12284	Same Context	Delanoue et al.,2004	14526388
Su(H)	CG3497	HLHmbeta	CG14548	Direct	Nellesen et al., 1999	11301266
Su(H)	CG3497	HLHm5	CG6096	Direct	Lecourtois & Schweisguth, 1995	7590238
Su(H)	CG3497	m2	CG6104	Direct	Lai et al., 2000	11301266
					Wakabayashi-Ito et al, 2001;	11133153
Su(H)	CG3497	fus	CG8205	Indirect	Doroquez & Rebay, 2006	17092823
Su(H)	CG3497	sim	CG7771	Direct	Morel & Schweisguth, 2000	10673509
					Bailey & Posakony, 1995;	7590239
Su(H)	CG3497	E(spl)	CG8365	Direct	Lecourtois & Schweisguth, 1995	7590238
gcm	CG12245	pnt	CG17077	Direct	Granderath et al., 2000	10704844
gcm	CG12245	repo	CG31240	Direct	Lee & Jones, 2005	15939231
gcm	CG12245	ttk	CG1856	Same Tissue	Badenhorst, 2001	11641231
gcm	CG12245	sim	CG7771	Same Tissue	Crews, 1992	1629656
h	CG6494	ac	CG3796	Direct	Van Doren et al., 1994	7958929
h	CG6494	ftz	CG2047	Direct	Carrol et al., 1988	3209072
twi	CG2956	meso18e	CG14233	Direct	Taylor, 2000	10720429
bcd	CG1034	gt	CG7952	Direct	Kraut & Levine, 1991	1893877

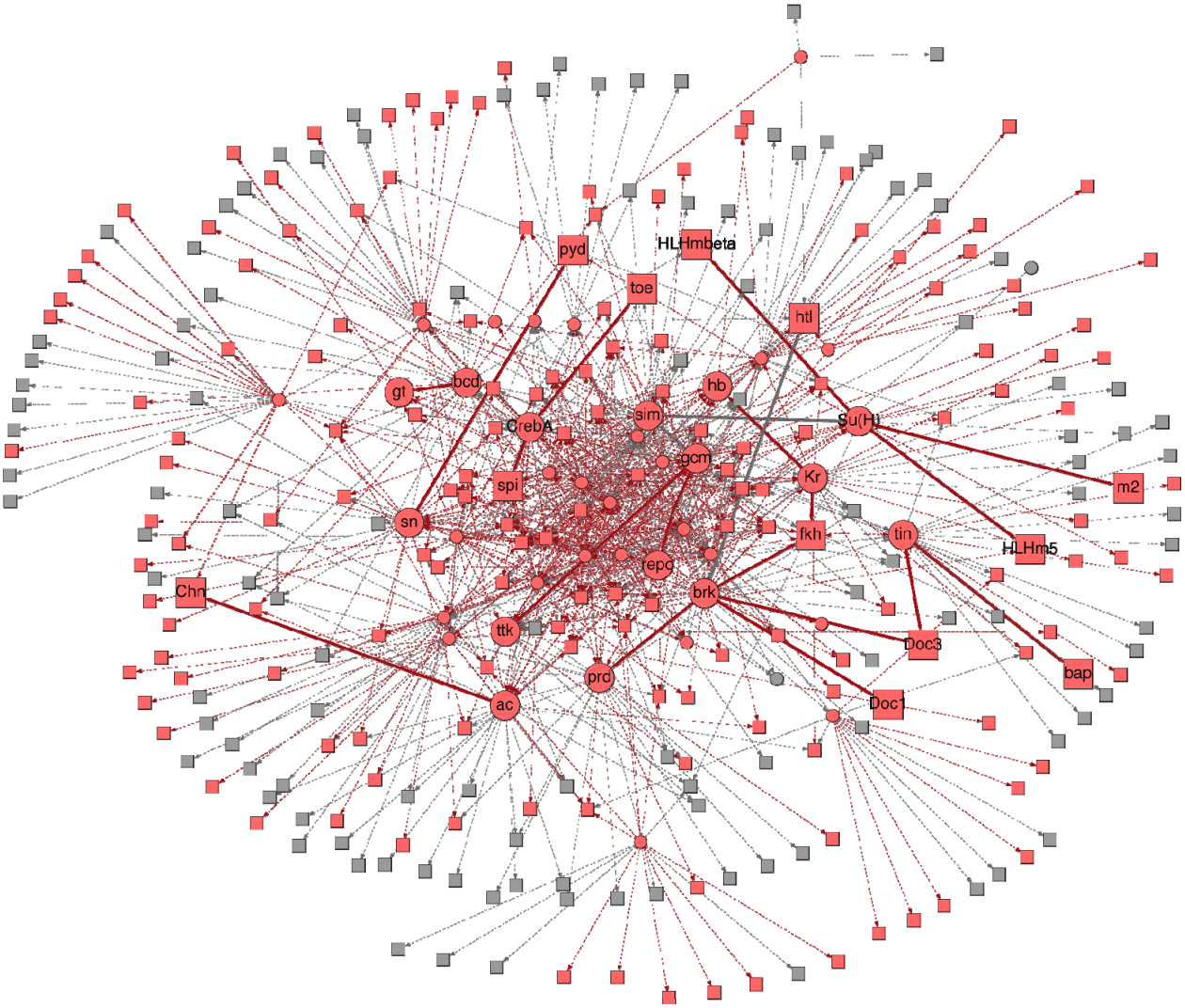
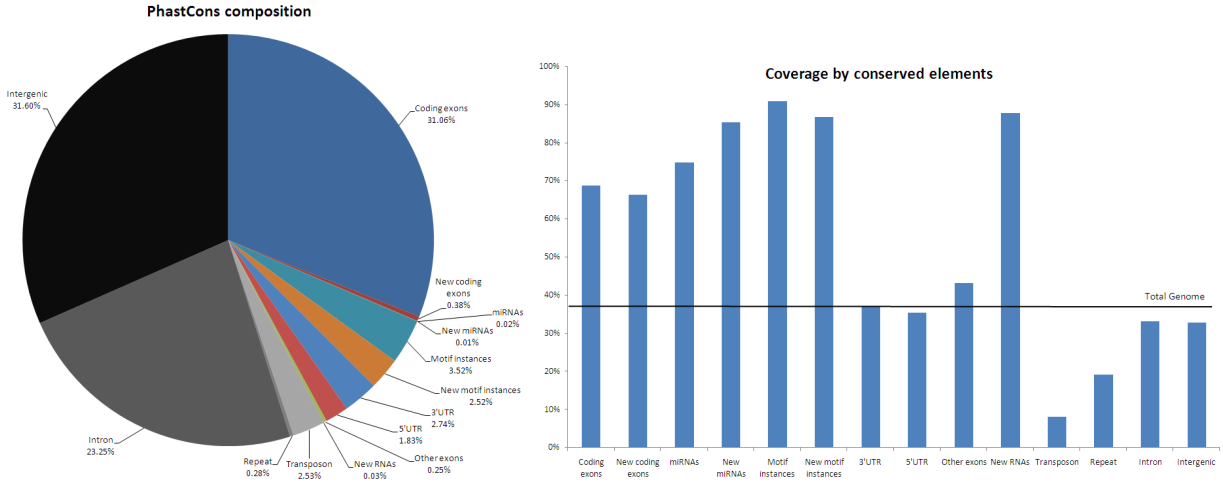


Figure S5i. Network of regulators and their targets identified at 60% confidence. Red nodes and edges indicate co-expression of factor (circles) and target (squares). Bold edges and gene names highlight known regulatory interactions (see Table S5f for a list of these).

S6 Comparison with phastCons elements



We measure the fraction of bases in phastCons elements covered by several annotations (left). Annotations include: known and new coding exons, cloned and predicted miRNAs, motif instances at 60% confidence for known and new motifs, 3' and 5' UTRs, other exons, new RNA structures, transposons, simple repeats, introns, and the remaining intergenic regions. When a region was annotated multiple categories, we assigned it to the first region category in the above list.

Known and predicted elements explain 42% of nucleotides in phastCons elements, leaving 3% as repeats/transposons, 23% in introns and 32% in unannotated intergenic space. This is an increase of 6.5% compared to annotated functional elements.

The chart on the right indicates the percentage of nucleotides covered by phastCons elements for each type of annotation.

S7 Scaling of comparative genomics power: additional information

S7a Protein-coding gene identification

To evaluate how multiple species can improve comparative gene identification, we assessed[8] the ability of our CSF metric to discriminate between known exons and random non-coding regions of similar lengths (Figure 8a). We found that multi-species comparisons consistently performed better than pairwise comparisons at similar evolutionary distances, especially for short exons. For example, at a fixed stringency (99% rejection of non-coding regions), the CSF metric applied to the best pairwise informant species recovered 88% of exons 40-60 amino acids in length, while with all 12 genomes, it recovered 94% of exons at the same stringency (a 50% reduction in false rejections). Over all sequence lengths, performance improved incrementally as more species were added at larger evolutionary distances, even though a relatively close species (*D. ananassae*) was the best overall pairwise informant.

nt	dere	dana	dpse	dwil	dgri
50	23.12%	45.36%	48.26%	**52.68%**	48.91%
110	62.98%	**85.91%**	84.35%	84.94%	83.68%
151	74.07%	**87.00%**	84.82%	86.57%	84.89%
196	82.43%	**88.46%**	85.26%	86.15%	84.88%
268	84.68%	**87.39%**	84.16%	84.09%	83.50%
388	**87.67%**	86.32%	84.98%	83.03%	81.69%
621	**91.08%**	89.60%	86.40%	84.99%	84.03%
1447	**93.24%**	91.76%	88.94%	87.82%	86.41%

Table S7a. Recovery rate for different pairwise informants in several categories (quantiles) of coding exon lengths, showing that closer species are preferred for longer exons.

S7b RNA structure prediction

We also found that increasing numbers of species led to increased power in ncRNA identification, based on the recovery of known ncRNAs among the top 100 predictions. The number of recovered ncRNAs increased roughly linearly with branch length (Figure 8b): the 5 *melanogaster* subgroup species recovered 14 ncRNAs, the 9 *Sophophora* species recovered 26, and all 12 species recovered 34, without apparent saturation.

S7c miRNA gene prediction

To assess contributions of additional species and increased evolutionary distance on miRNA discovery power, we studied our miRNA recovery rate among the top 100 predictions (Figure 8c). For pairwise comparisons, this rate increased with evolutionary distance: any of the five closest species led to the recovery of up to 49 miRNAs, any of the next four species up to 73, and any of the three most distant species up to 76. Additionally, multi-species comparisons led to increased power over pairwise comparisons at similar evolutionary distances, recovering 49, 78, and 84 miRNAs using 5, 9, and 12 species, respectively.

We also asked whether using subsets of the species would allow us to discover any of the 32 lineage-specific miRNAs that were not found using all 12 species. However, we recovered only one of these miRNAs among 21 different pairwise or multi-species comparisons (among the same total number of predicted hairpins), and the recovery using all 12 species was always best overall. This suggests that the currently sequenced species do not provide sufficient power to identify these more recently evolved miRNAs and perhaps that their conservation patterns or structural properties differ from those in our training set (Rfam miRNAs).

S7d Motif instance prediction

Finally, we investigated how multiple informant species enhance the identification of individual motif instances. We first compared the conservation of known motifs to that of randomly shuffled control motifs (signal-to-noise ratio, SNR) with different species subsets (Figure 8d). We found that for both transcription factor (TF) and miRNA motifs, the average SNR increased with the addition of more species at larger evolutionary distances. For example, from 6 to 12 species, the average SNR increased from 2:1 to 3:1 for TF motifs, and from 2.5:1 to 8:1 for miRNA motifs.

Additional species also allowed us to recover more high-confidence motif instances. First, across 12 species, more TF and miRNA motifs reached confidence levels of 60% or above (for at least one instance), than for subsets of these species (see below). Additionally, our BLS conservation measure that tolerates motif loss allowed much higher sensitivity than simpler strategies of requiring perfect conservation across 12, 9, 5, or 2 species, and recovered more experimentally supported motif instances[42] (Figure 8f).

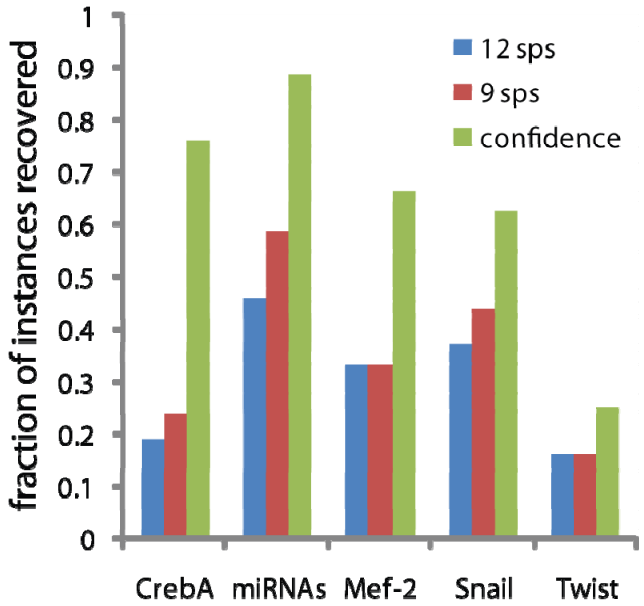


Figure S7d. Comparison of sensitivity in recovering validated functional instances using a confidence (evaluated via BLS) versus perfect conservation in 9 or 12 species, which enable sufficient specificity. The improved sensitivity of the BLS holds for all genome-wide instances for all known motifs at all confidence cutoffs[42].

S8. Influence of alignments on comparative predictions

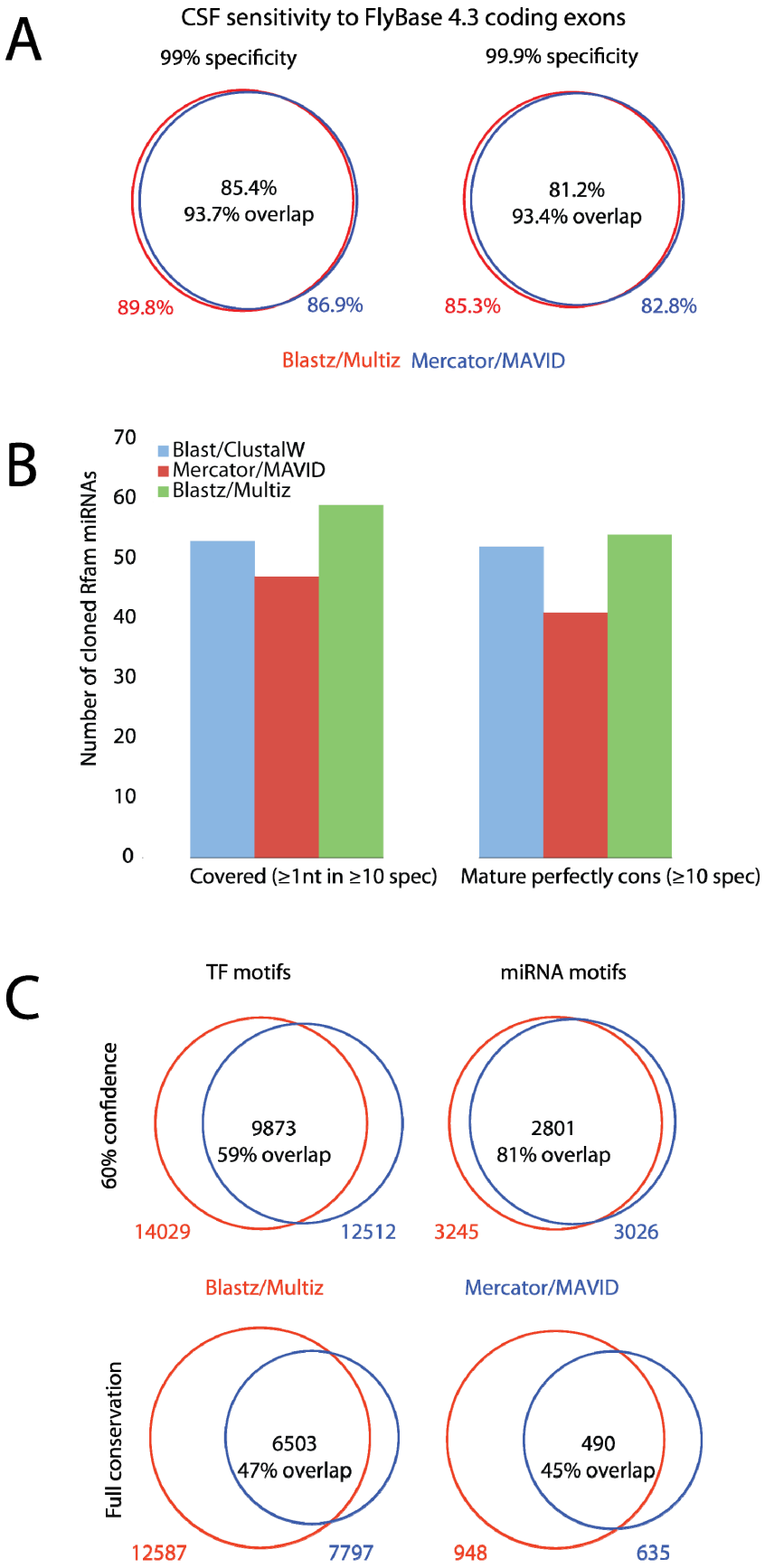
To gauge the influence of genome alignments on the analyses presented in this paper, we applied our methods to different alignments and compared the results (see next page).

A. **Protein-coding exons.** Proportion of FlyBase 4.3 exons that are distinguishable from random non-coding regions using CSF, at a fixed specificity in each genome alignment. Overall, the two genome alignments led to highly concordant results for coding exons: at 99% specificity, 94% of the exons detected in either alignment were detected in both alignments, with the BlastZ/MultiZ alignments leading to somewhat higher overall sensitivity (90% vs. 87%). We observed similar trends at a more stringent cutoff, as well as with RFC and other metrics[8]. These data suggest that the BlastZ/MultiZ alignments might have allowed slightly more (~3%) new exon predictions at similar predictive value.

B. **miRNA genes.** Compared to the Mercator/MAVID alignment (the only whole-genome alignment available at the time of this part of the study), our Blast/ClustalW alignments had sequence coverage across ≥10 species for more miRNA hairpins, and revealed perfect conservation across ≥10 species for more mature miRNAs. On the other hand, the more recently produced Blastz/Multiz alignment shows a slightly higher sensitivity than our Blast/ClustalW alignments, covering 59 (vs. 53) miRNA hairpins and showing perfect conservation for 54 (vs. 52) mature miRNAs. The discrepancy suggests that our predictions might have missed a small fraction of miRNAs due to alignment issues.

C. **Motif instances.** As expected given their short lengths, prediction of motif instances was influenced by alignment choice more than protein-coding exons or miRNAs. While the majority of predicted instances were found using both genome-wide alignments (81% of miRNA and 59% of TF motif instances), alignment discrepancies indicated that the true number of conserved instances might be up to 7% higher for miRNA motifs and 19% higher for TF motifs. Notably, the agreement between the two genome alignments is higher for motif instances identified at 60% confidence based on BLS (81% of miRNA motif instances and 59% of TF motif instances in promoters) than when requiring perfect conservation across all 12 species (45% and 47%, respectively), suggesting the importance of methods that account for potential alignment discrepancies. We found that the Blastz/Multiz alignments consistently identified more motif instances above noise than the Mercator/MAVID alignments (7% more miRNA motif instances and 11% more TF motif instances at 60% confidence). This is unlikely to be due to lower specificity of the Blastz/Multiz alignments, as our confidence measure corrects for increased overall sequence conservation/similarity, and the increased recovery of instances indicates that that more orthologous regulatory sequences have been aligned correctly.

We conclude that the different alignment strategies agree for the vast majority of elements predicted in our analysis. We find somewhat higher overall sensitivity when using the Blastz/Multiz alignments, but the Mercator/MAVID alignment also captures elements not recovered by Blastz/Multiz. We note that signal-to-noise measures or motif-recovery-over-noise used here might constitute an attractive means to assess alignment quality and coverage in intergenic regions, which have traditionally been more difficult to align and assess. With ongoing efforts to compare, evaluate and improve genome-wide alignments, comparative predictions are likely to become increasingly consistent across alignments[66,67].


A
CSF sensitivity to FlyBase 4.3 coding exons
99% specificity
99.9% specificity
85.4%
93.7% overlap
89.8%
86.9%
81.2%
93.4% overlap
85.3%
82.8%
Blastz/Multiz
Mercator/MAVID
B
Blast/ClustalW
Mercator/MAVID
Blastz/Multiz
Number of cloned Rfam miRNAs
70
60
50
40
30
20
10
0
Covered (≥1nt in ≥10 spec)
Mature perfectly cons (≥10 spec)
C
TF motifs
miRNA motifs
60% confidence
9873
59% overlap
14029
12512
2801
81% overlap
3245
3026
Blastz/Multiz
Mercator/MAVID
Full conservation
6503
47% overlap
12587
7797
490
45% overlap
948
635

S9 Data availability and accession numbers

In addition to the information provided in this supplement, we provide online supplemental materials including large lists, tables, and datasets.

Genome sequence alignments

Mercator/MAVID	http://www.biostat.wisc.edu/~cdewey/fly_CAF1/
Mercator/PECAN	http://www.sanger.ac.uk/Users/td2/pecan-CAF1/
MULTIZ	http://hgdownload.cse.ucsc.edu/goldenPath/dm2/multiz15way/

Revisiting the protein-coding gene catalog

Tables of scores for all existing genes, list of "rejected" genes, proposed corrections to existing annotations, new gene predictions, FlyBase curation records, cDNA validation sequences, GenBank accession nos., genes with unusual features, etc. are available in the online supplement for:

Lin, Carlson, Crosby *et al.* (2007). Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using twelve fly genomes. *Genome Res.*, in press.

http://www.broad.mit.edu/~mlin/fly_genes/

Full-length cDNA sequences were deposited into GenBank, accession numbers BT029554…BT029635, BT029637…BT029727, BT029940…BT029957, BT030133…BT030144, BT030416…BT030421, and BT030448…BT030452.

RNA gene prediction

Tables as well as files defining both the comprehensive and the high-confidence prediction sets can be found at: http://www.soe.ucsc.edu/~jsp/flyFolds/

miRNA gene prediction

All data for miRNA gene prediction are available in the online supplement for:

Stark, Kheradpour, Parts, Brennecke, Hodges, Hannon, Kellis (2007). Systematic discovery and characterization of fly miRNAs using 12 *Drosophila* genomes. *Genome Res.*, in press.

http://compbio.mit.edu/fly/mirnas/

Small RNA sequences[40] were deposited into the Gene Expression Omnibus, accession numbers GPL5061 and GSE7448.

miRNA target prediction

All miRNA target predictions are available at http://www.targetscan.org/

Regulatory motif prediction

Novel motifs and their properties and all regulatory target predictions at different confidence levels are available at http://compbio.mit.edu/fly/regulation/

S10 References for supplementary materials

1 Bray, N. & Pachter, L. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* 14 (4), 693-699 (2004).

2 Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14 (4), 708-715 (2004).

3 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* 52 (5), 696-704 (2003).

4 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5 (1), 113 (2004).

5 Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* 12 (6), 996-1006 (2002).

6 Kellis, M. *et al.* Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol* 11 (2-3), 319-355 (2004).

7 Kellis, M. *et al.* Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423 (6937), 241-254 (2003).

8 Lin, M.F., Deoras, A. N., Rasmussen, M. D., & Kellis, M. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *Genome Res* submitted (2007).

9 Lafferty, J., McCallum, A., & Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th International Conf. on Machine Learning*, 282–289 (2001).

10 Sarawagi, S. & Cohen, W. Semi-markov conditional random fields for information extraction. *Advances in Neural Information Processing Systems* 17, 1185-1192 (2005).

11 Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268 (1), 78-94 (1997).

12 Gross, S. S. & Brent, M. R. Using multiple alignments to improve gene prediction. *J Comput Biol* 13 (2), 379-393 (2006).

13 Decaprio, D. *et al.* Conrad: Gene prediction using conditional random fields. *Genome Res* 17 (9), 1389-1398 (2007).

14 Bernal, Axel E., Crammer, Koby, Hatzigeorgiou, Artemis, & Pereira, Fernando C. N. Global Discriminative Training for Higher-Accuracy Computational Gene Prediction. *PLoS Computational Biology* preprint (2007), e54.eor (2007).

15 Siepel, A. & Haussler, D. presented at the Proc 8th Annual Int'l Conf on Research in Computational Biology RECOMB'04, 2004 (unpublished).

16 Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 11 (2-3), 377-394 (2004).

17 Pedersen, J. S. *et al.* Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput Biol* 2 (4), e33 (2006).

18 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15 (8), 1034-1050 (2005).

19 Washietl, S. *et al.* Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* 17 (6), 852-864 (2007).

20 Stapleton, M., Carlson, J. W., & Celniker, S. E. RNA editing in Drosophila melanogaster: New targets and functional consequences. *Rna* 12 (11), 1922-1932 (2006).

21 Howard, M. T. *et al.* Recoding elements located adjacent to a subset of eukaryal selenocysteine-specifying UGA codons. *Embo J* 24 (8), 1596-1607 (2005).

22 Stefanovic, B. & Brenner, D. A. 5' stem-loop of collagen alpha 1(I) mRNA inhibits translation in vitro but is required for triple helical collagen synthesis in vivo. *J Biol Chem* 278 (2), 927-933 (2003).

23 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25 (1), 25-29 (2000).

24 Manak, J. R. *et al.* Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat Genet* 38 (10), 1151-1158 (2006).

25 Kuhn, R. M. *et al.* The UCSC genome browser database: update 2007. *Nucleic Acids Res* 35 (Database issue), D668-673 (2007).

26 Griffiths-Jones, S. *et al.* miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 34 (Database issue), D140-144 (2006).

27 Cohen, R. S., Zhang, S., & Dollar, G. L. The positional, structural, and sequence requirements of the Drosophila TLS RNA localization element. *Rna* 11 (7), 1017-1029 (2005).

28 Griffiths-Jones, S. *et al.* Rfam: an RNA family database. *Nucleic Acids Res* 31 (1), 439-441 (2003).

29 Mignone, F. *et al.* UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res* 33 (Database issue), D141-146 (2005).

30 Xia, S. *et al.* Identification of new targets of Drosophila pre-mRNA adenosine deaminase. *Physiol Genomics* 20 (2), 195-202 (2005).

31 Diegelmann, S. *et al.* The conserved protein kinase-A target motif in synapsin of Drosophila is effectively modified by pre-mRNA editing. *BMC neuroscience* 7, 76 (2006).

32 Hoopengardner, B., Bhalla, T., Staber, C., & Reenan, R. Nervous system targets of RNA editing identified by comparative genomics. *Science* 301 (5634), 832-836 (2003).

33 Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly* V125 (2), 167 (1994).

34 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25 (17), 3389-3402 (1997).

35 Thompson, J. D., Higgins, D. G., & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22 (22), 4673-4680 (1994).

36 Breiman, L. Random Forests. *Machine Learning* 45, 5-32 (2001).

37 Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434 (7031), 338-345 (2005).

38 Stark, A. *et al.* Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3'UTR Evolution. *Cell* 123 (6), 1133-1146 (2005).

39 Joachims, T. in *Advances in Kernel Methods - Support Vector Learning*, edited by B. Schölkopf, C. Burges, & A. Smola (Cambridge, Mass., 1999), pp. 41-56.

40 Ruby, J. G. *et al.* Evolution, Biogenesis, Expression, and Target Predictions of a Substantially Expanded Set of Drosophila MicroRNAs. *Genome Res* in press (2007).

41 Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* 128 (6), 1089-1103 (2007).

42 Stark, A., Kheradpour, P., Roy, S., & Kellis, M. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res* submitted (2007).

43 Felsenstein, Joseph *Inferring phylogenies*. (Sinauer Associates, Sunderland, Mass., 2004).

44 van Helden, J., Rios, A. F., & Collado-Vides, J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* 28 (8), 1808-1818 (2000).

45 Agresti, A. & Coull, B. A. Approximate is better than 'exact' for interval estimation of binomial parameters. *The American Statistician* 52, 119-126 (1998).

46 Brown, L. D., Cai, T. T., & DasGupta, A. Interval estimation for a binomial proportion. *Statist. Sci* 16, 101-133 (2001).

47 Dewey, C. N. *et al.* Parametric alignment of Drosophila genomes. *PLoS Comput Biol* 2 (6), e73 (2006).

48 Lewis, B. P. *et al.* Prediction of mammalian microRNA targets. *Cell* 115 (7), 787-798 (2003).

49 Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 31 (1), 374-378 (2003).

50 Sandelin, A. *et al.* JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32 (Database issue), D91-94 (2004).

51 Bergman, C. M., Carlson, J. W., & Celniker, S. E. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics* 21 (8), 1747-1749 (2005).

52 Pietrokovski, S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res* 24 (19), 3836-3845 (1996).

53 Schones, D. E., Sumazin, P., & Zhang, M. Q. Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics* 21 (3), 307-313 (2005).

54 Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. Quantifying similarity between motifs. *Genome Biol* 8 (2), R24 (2007).

55 Sethupathy, P., Corda, B., & Hatzigeorgiou, A. G. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *Rna* 12 (2), 192-197 (2006).

56 Sandmann, T. *et al.* A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes Dev* 21 (4), 436-449 (2007).

57 Sandmann, T. *et al.* A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev Cell* 10 (6), 797-807 (2006).

58 Zeitlinger, J. *et al.* Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes Dev* 21 (4), 385-390 (2007).

59 Abrams, E. W. & Andrew, D. J. CrebA regulates secretory activity in the Drosophila salivary gland and epidermis. *Development* 132 (12), 2743-2758 (2005).

60 Stanley, S. M., Bailey, T. L., & Mattick, J. S. GONOME: measuring correlations between GO terms and genomic positions. *BMC Bioinformatics* 7, 94 (2006).

61 Bollobás, Béla *Random graphs*, 2nd ed. (Cambridge University Press, Cambridge ; New York, 2001).

62 Adryan, B. & Teichmann, S. A. FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. *Bioinformatics* 22 (12), 1532-1533 (2006).

63 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13 (11), 2498-2504 (2003).

64 Farh, K. K. *et al.* The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* 310 (5755), 1817-1821 (2005).

65 Sood, P. *et al.* Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc Natl Acad Sci U S A* (2006).

66 Margulies, E. H. *et al.* Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* 17 (6), 760-774 (2007).

67 Prakash, A. & Tompa, M. Measuring the accuracy of genome-size multiple alignments. *Genome Biol* 8 (6), R124 (2007).